РОССИЙСКАЯ АКАДЕМИЯ НАУК
Сибирское отделение РАН

ИНСТИТУТ ОРГАНИЧЕСКОЙ ХИМИИ

ОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В.И. ВЕРШИНИН, Б.Г. ДЕРЕНДЯЕВ, К.С. ЛЕБЕДЕВ

КОМПЬЮТЕРНАЯ ИДЕНТИФИКАЦИЯ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

«НАУКА»

2002

УДК 543.42

Компьютерная идентификация органических соединений. В е р ш и н и н В. И., Д е р е н д я е в Б. Г., Л е б е д е в К. С. М.: Наука. 2002.

Монография посвящена новому аналитическому методу – опознанию органических соединений с помощью ЭВМ на основе банков данных, включающих спектральные, масс-спектрометрические и хроматографические характеристики индивидуальных соединений. Описаны общие принципы, алгоритмы, критерии, результаты и практическое применение метода. Основное внимание уделено наиболее сложным проблемам: установлению молекулярной структуры ранее неизвестных соединений и качественному анализу многокомпонентных смесей.

Книга предназначена для химиков-аналитиков, химиков-органиков, специалистов в области хроматографии и спектрального анализа, а также математиков и физиков, интересующихся компьютерными методами идентификации веществ. Поскольку книги по данной тематике ранее в России не издавались, монография может быть полезна и в качестве учебного пособия для студентов-химиков старших курсов и аспирантов.

С Издательство «Наука», 2002 г.

Оглавление

	Стр.
Предисловие	3
Введение	5
Глава 1. Общие принципы компьютерной идентификации индивидуальных соединений на основе баз данных. Опознание известных соединений с помощью информационно-поисковых систем	8
1.1.. Базы данных, используемые при решении идентификационных задач в различных методах анализа.	8

1.2. Информационно-поисковые системы для опознания соединения по его спектру	32
Глава 2. Компьютерные средства и методы установления строения неизвестного соединения по спектральным данным	57
2.1. Методология решения структурных задач с помощью баз данных	58
2.2. Информационно-аналитические системы по молекулярной спектроскопии	94
Глава 3. Компьютерный качественный анализ смесей по характеристикам удерживания при хроматографическом разделении. Проблема идентификации компонентов с заданной надежностью	102
3.1. Качественный хроматографический анализ по характеристикам удерживания	102
3.2. Компьютерные ИПС в хроматографическом анализе	107
3.3. Вероятностная оценка надежности идентификации	110
3.4. Выбор критериев идентификации	119
3.5. Хроматографическая идентификация при многократных испытаниях	123
Глава 4. Качественный спектральный анализ неразделенных смесей.	129
4.1. Общие подходы к качественному анализу смесей.	129
4.2. Методология обратного поиска в анализе смесей.	133
4.3. Алгоритмы анализа смесей с применением вероятностных критериев	136
4.4. Применение вероятностных критериев при расшифровке спектров НЛ. Оценка критериев в ходе компьютерных экспериментов	145
4.5. Учет специфичности и относительной интенсивности линий	152
4.6. Анализ модельных и реальных смесей полиаренов с помощью ИПС “Спектр”	156
Заключение	164
Литература	169

Предисловие

Аналитическая химия начиналась как наука об анализе неорганических веществ, и долгие десятилетия, если не столетия, и в своем теоретическом базисе, и в практических методиках основным объектом приложения сил она видела минеральное сырье, металлы и другие объекты неорганической природы. Еще и сейчас такой подход сквозит в некоторых учебных программах и учебниках по аналитической химии. Но времена меняются, сейчас органические вещества и близкие к ним биообъекты заняли подобающее им место; многие современные методы химического анализа и нацелены главным образом на такие объекты.

Качественный и количественный анализы развивались параллельно, но не всегда одинаково интенсивно. Пока атомно-эмиссионный, рентгенофлуоресцентный анализ и другие физические методы, легко решающие задачи и качественного и количественного анализа, не стали обычными, качественным анализом чаще всего считали обнаружение, главным образом химическими методами, катионов, анионов и реже (в фазовом анализе) отдельных неорганических соединений. Хотя реакции обнаружения некоторых органических соединений и особенно функциональных групп, а следовательно, классов веществ, стали появляться еще в середине XIX века, качественный анализ веществ органической природы длительное время не был предметом широкого внимания. Однако резкое расширение номенклатуры известных органических соединений, широчайшее их использование, осознание необходимости находить опасные вещества и т.д. вызвали к жизни разработку методов их обнаружения, идентификации, ну и, конечно, количественного определения. Среди методов прежде всего хроматографические, спектроскопические, масс-спектромет-рические, электрохимические и др.

Качественный анализ органических веществ сейчас имеет огромное значение. Идентификация компонентов сложных смесей, например, при анализе нефтепродуктов или объектов окружающей среды; обнаружение наркотиков, взрывчатых и отравляющих веществ, оценка подлинности лекарств, установление природы вещества, синтезированного в процессе научной работы, - все это малая часть перечня направлений и проблем, требующих качественного органического анализа.

Идентификация органических соединений, особенно со сложной формулой,- дело непростое и трудоемкое. Особенно это относится к так называемым неизвестным соединениям, которые не гарантированно индивидуальными, без примесей, а, напротив, присутствуют в составе сложных смесей. Сколько работ посвящено, например, идентификации соединений по характеристикам удерживания в хроматографии, а проблема до конца не решена. В наше время естественным было стремление использовать компьютеры, причем почти сразу наметились два пути: 1) создание баз данных и информационно-поисковых систем и 2) разработка систем искусственного интеллекта, в настоящее время получивших название экспертных систем. Оба подхода успешно развиваются и в нашей стране, причем первый – активнее всего сибирскими коллегами, по инициативе и продолжительное время под руководством академика В.А.Коптюга.

Эта книга обобщает прежде всего работу этой группы. Для компьютерной идентификации органических соединений используются базы данных масс-спектрометрии, ИК спектроскопии и ядерного магнитного резонанса (^{1}H и ^{13}C) и соответствующие информационно-поисковые системы. Сделанное обобщение – глубокое и полное, с широким привлечением мировой литературы по этому вопросу.

Однако, авторы не ограничились лишь использованием спектроскопических и масс-спектрометрических данных (несмотря на схожесть названий, масс-спектрометрия – не спектроскопический метод). Много внимания уделено компьютерной идентификации с применением параметров хроматографического удерживания. Это качественный анализ сложных смесей. И еще: последняя глава посвящена качественному спектроскопическому анализу неразделенных смесей.

Далее требуется употребить словосочетание, которое у кого-то вызовет недоумение: метрология качественного анализа. (На самом деле такой метрологией Н.П.Комарь занимался в Харькове еще в 50-60 гг.) В этой книге много сказано о количественной оценке надежности идентификации. Разве это не метрология качественного анализа?

Авторы монографии (кстати, кажется, первой на данную тему) – известные специалисты своего дела. Доктор химических наук В.И.Вершинин заведует кафедрой аналитической химии и химии нефти Омского университета, крупный ученый в области анализа смесей органических веществ, в том числе методом хроматографии. Доктора химических наук Б.Г.Дерендяев и К.С.Лебедев – ведущие сотрудники покойного В.А.Коптюга, они из Научно – технического центра химической информатики при Новосибирском институте органической химии СО РАН (К.С.Лебедев сейчас работает в Новомосковске).

Монография прекрасно написана; построена логично, материал охвачен широко, иллюстрации (понятные даже неспециалисту) – по делу и к месту. Но что хотелось бы подчеркнуть особо, так это отличный стиль; редко бывает, что при чтении научной книги не спотыкаешься на выкрутасах русского языка. Причем я читал рукопись еще до ее шлифовки издательским редактором.

Адресаты в книге указаны. Указаны, я думаю, правильно.

Академик Ю.А. Золотов

ВВЕДЕНИЕ

Стремительное развитие средств вычислительной техники, появление автоматизированных систем регистрации и обработки сигналов в сочетании с возможностью использования крупных баз данных привели в конце XX века к рождению нового раздела аналитической химии. Речь идет об *аналитической химии, основанной на использовании компьютеров* [1]. В англоязычной литературе принято сокращенное наименование COBAC (computer based analytical chemistry). Являясь одновременно частью нового направления химической науки в целом (*компьютерной химии*) и широко используя алгоритмы хемометрики, методы COBAC призваны обеспечить машинную переработку совокупности аналитических сигналов для получения достоверной качественной и количественной информации, характеризующей исследуемый материал. Важной частью этого научного направления стало применение компьютерных технологий в качественном анализе.

В течение длительного времени качественный анализ развивался намного медленнее количественного. Ряд монографий и учебников до сих пор уделяют основное внимание малозначимому для современной аналитической службы частному случаю - обнаружению некоторых катионов металлов с помощью химических реакций в растворе [2]. Литература по инструментальным методам преимущественно посвящена опознанию чистых веществ, синтезированных в лаборатории или хроматографически выделенных из природной смеси. Однако традиционные способы идентификации органических соединений (по атласам, таблицам и т.п.) весьма трудоемки и длительны, причем основные трудности связаны не с самим получением спектра или хроматограммы пробы, а с надежной интерпретацией

полученных данных [3]. Невозможно вручную сопоставить спектр пробы с десятками (а то сотнями) тысяч эталонных спектров разных органических соединений и отыскать среди них наиболее похожие. Поэтому новейшие методики предусматривают использование компьютеров как для управления приборами и регистрации полученных данных (спектров, хроматограмм и т.п.), так и для логической интерпретации этих данных на основании информации, найденной в соответствующих базах данных. Компьютерный поиск информации проводится и при опознании известных веществ, и при исследовании строения впервые синтезированных органических соединений, и в анализе неразделенных многокомпонентных смесей.

Базы данных (БД) и работающие с ними компьютерные информационно-поисковые системы (ИПС) стали применяться аналитиками еще в конце 60-х годов; сначала в рентгенофазовом анализе, инфракрасной и масс-спектрометрии, а затем и в других спектроскопических, хроматографических, резонансных и гибридных методах. Число различных ИПС, нацеленных на опознание чистых веществ, теперь измеряется сотнями. Огромные по своему объему БД, а также ИПС универсального характера, входящие в состав национальных и международных информационных систем обеспечивают доступ (в том числе удаленным пользователям) к любой химической информации. Небольшие по объему и, как правило, проблемно-ориентированные БД и ИПС входят в состав программного обеспечения спектрометров и хроматографов высшего класса. Они приобрели самостоятельную коммерческую ценность, стали объектом патентования. Сертифицированные методики компьютерного качественного анализа включаются в сборники нормативных документов (ASTM и др.). Соответствующие приборы, программы и методики вошли в практику многих российских лабораторий - исследовательских и даже заводских. Однако статьи, описывающие алгоритмы и результаты компьютерного качественного анализа, в последние годы стали реже появляться в научной печати: информация о конкретных программах теперь дается производителями с учетом соображений маркетинга и коммерческой тайны.

К сожалению, отечественные монографии и учебники по аналитической химии практически не затрагивают вопросы, связанные с компьютеризацией качественного анализа (исключением является давняя монография [4]). В то же время несомненно, что в последние десятилетия в этой области достигнуты результаты, влияющие на методологию и метрологию химического анализа в целом. В частности, создание автоматизированных систем для идентификации органических соединений потребовало более высокого уровня теоретических обобщений, разработки формализованных алгоритмов принятия решений, применения количественных и метрологически обоснованных оценок, развития

соответствующих математических методов. Двумя основными и дополняющими друг друга подходами к компьютерному качественному анализу являются: использование метода искусственного интеллекта [4] и применение фактографических баз данных. Отметим, что для опознания органических соединений с успехом используют искусственные нейронные сети и методы распознавания образов, но соответствующие алгоритмы можно реализовать в рамках того или другого из вышеуказанных основных подходов.

Научная терминология в данной области еще не полностью устоялась. Поэтому следует уточнить, что термином “*идентификация*” мы называем не процесс, а результат проведенного качественного анализа, логический вывод, а именно - признание некоторого компонента присутствующим в исследуемой пробе (естественно, с определенной вероятностью ошибки). В этом же смысле термин определяет и "Химическая энциклопедия". Для заведомо однокомпонентных проб “идентификация” означает признание тождественности пробы и вещества-эталона. Такое определение - частный случай первого, более общего. Сочетание термина “идентификация” с прилагательным “*компьютерная*” подчеркивает роль и значимость средств вычислительной техники, информационных технологий и баз данных в принятии решения о составе пробы. Эти средства позволяют не только умножить возможности аналитика и используемых им приборов, но и получить качественно новые сведения об объекте анализа с одновременной оценкой их достоверности. Другие термины и обозначения будут вводиться в книге по ходу изложения материала.

Из многочисленных и разнообразных инструментальных методов, применяемых для решения идентификационных задач, нами выделены наиболее эффективные – молекулярная спектроскопия и хроматография. Авторы стремились продемонстрировать достижения в рассматриваемой области, привлечь внимание аналитиков к необходимости действенного участия в создании новых и наполнении существующих баз данных, показать перспективы их использования в качественном анализе, в том числе для решения такой актуальной задачи как опознание веществ с заданной надежностью [5].

Методу компьютерной идентификации веществ с применением БД и ИПС посвящены вышедшие на английском языке книги Мак-Лафферти, Зупана, Грея и других авторов [6-9]. Российские ученые также внесли и продолжают вносить весомый вклад в развитие теории и практики компьютерной идентификации органических соединений (см. обзоры в [10-14]). Следует, в частности, отметить заслуги акад. В.А.Коптюга, создавшего еще в 70-ые годы новосибирскую школу исследователей - на стыке аналитической химии, спектроскопии, информатики и структурной органической химии. Однако на русском языке книги по компьютерной идентификации органических соединений с применением БД и ИПС до сих пор не издавались. Именно в этой области работают авторы, и в книге изложены

результаты исследований соответствующих научных коллективов. Глава 1 написана д.ф.м..н. Б.Г. Дерендяевым, глава 2 - д.х.н. К.С. Лебедевым, главы 3 и 4 - д.х.н. В.И. Вершининым. Последний автор взял на себя и общую редакцию книги.

Мы глубоко благодарны академику Ю.А.Золотову, любезно согласившемуся написать краткое предисловие к этой книге и давшему ряд важных советов по ее содержанию. Выражаем свою признательность коллегам, участововавшим в исследованиях, результаты которых позволили написать эту книгу. Мы будем рады всем замечаниям и предложениям по содержанию книги и по обсуждаемым в ней проблемам. Электронные адреса авторов:
vershin@univer.omsk.su der@nioch.nsc.ru Analitika@ novomoskovsk.ru

Монография написана в год семидесятилетия академика В.А.Коптюга, стоявшего у истоков компьютерной химии и химической информатики. Его светлой памяти авторы и посвящают эту книгу.

Глава 1
ОБЩИЕ ПРИНЦИПЫ КОМПЬЮТЕРНОЙ ИДЕНТИФИКАЦИИ ИНДИВИДУАЛЬНЫХ СОЕДИНЕНИЙ НА ОСНОВЕ БАЗ ДАННЫХ. ОПОЗНАНИЕ ИЗВЕСТНЫХ СОЕДИНЕНИЙ С ПОМОЩЬЮ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

1.1. Базы данных, используемые в различных методах анализа

1.1.1. Общие сведения о химических базах данных

Наиболее полные и общие сведения об изучаемом химическом объекте, в частности, о его физико-химических свойствах, могут быть найдены в крупных *банках данных коллективного пользования*. Соответствующие БД сегодня доступны большинству потребителей информации, в том числе по телекоммуникационным линиям связи в режиме on-line. В России основными источниками являются БД Всероссийского института научной и технической информации (ВИНИТИ) и БД Международной научно-технической сети STN International.

В политематической БД ВИНИТИ реферируется и сохраняется около 1 млн. публикаций в год. Доступ к ее “химическому” разделу возможен по сети Интернет

(www.viniti.ru) или путем приобретения (закупки) БД на дисках CD. Подробное знакомство с методами и приемами поиска информации в доступных по сети INTERNET базах данных ВИНИТИ возможно с помощью специальных изданий (например [15,16]).

STN International объединяет материалы трех крупнейших производителей вторичной научной информации: Chemical Abstracts Service (CAS), Fachinformationszentrum Karlsruhe (FIZ K) и Japan Science and Technology Corporation (JST). Все они соединены спутниковой и кабельной связью, а пользователь получает доступ ко всем БД STN International, присоединившись к любому из указанных центров. Общее количество доступных в STN баз данных около 200.

CAS создает группу баз данных по различным областям знания, примыкающим к области химии и объединенных в CAS Online. Например, в базе данных Registry зарегистрировано с 1957 года свыше 28,7 миллионов (на 18.01.2001 г.) веществ (сплавов, составов, композиций), сведения о которых опубликованы в мировой литературе. FIZ K специализируется в областях физики, энергетики, патентов, вычислительной техники и т.п. JST – ведет англоязычные БД по японской научно-технической литературе. STN обеспечивает доступ не только к БД собственного производства, но и к целому ряду других, адаптированных к проведению поиска на едином языке, принятом в STN. Ниже приводятся краткие сведения об основных “химических” БД сети STN International.

База данных **Chemical Abstracts (CA)** содержит рефераты публикаций по химии и охватывает области химии, химической технологии, биохимии и смежных дисциплин с 1967 года по настоящее время. Источниками информации являются более 9 тыс. научных журналов, книги, обзоры, патентные документы, труды конференций и т.п. Тематика CA включает органическую, неорганическую, физическую, аналитическую, прикладную и макромолекулярную химию.

Файл **BEILSTEIN** содержит информацию об органических соединениях, методах их синтеза и реакциях, а также справочные данные о свойствах веществ. Источники информации этой БД – “Справочник Бейльштейна по органической химии” (Beilstein Handbuch der Organishe Chemie) и данные из 140 ведущих журналов по органической химии с 1779 года по настоящее время. Поиск информации можно проводить по структурным формулам, регистрационным номерам, физико-химическим данным и другим свойствам. Аналогичные виды поисков реализуются в CA.

GMELIN - файл по неорганической и металлорганической химии. Он содержит тщательно отобранные численные данные из справочника “Gmelin Handbook of Inorganic and Organometallic Chemistry” (с 1817), а также избранные данные из 112 важнейших журналов по неорганической, физической и металлорганической химии с 1975 по 1997 гг. В этом

файле возможен поиск по библиографической информации, регистрационным номерам, химическим названиям, молекулярным формулам, структурам и численным данным химических и/или физических свойств веществ.

HSDB – содержит информацию по токсичности и экологической безопасности химических веществ. Записи БД включают сведения о методах обнаружения веществ, воздействии на среду и организм человека, нормативные документы, сведения о производстве и использовании химических веществ. Фактографические данные базы: название соединения, структурная формула, регистрационный номер CAS, стандарты, нормативы, физические и химические свойства.

MRCK (или **MERCK**) – содержит краткое описание лекарственных препаратов, биологических и природных продуктов, сельскохозяйственных химикатов, в том числе и применяемых в ветеринарии, органических и неорганических веществ, используемых в исследовательских и коммерческих целях. Записи содержат названия соединений, регистрационный номер CAS, коды лекарств, торговые названия, сведения о производителях, физические и токсикологические данные, библиографические ссылки на научную и патентную литературу.

HODOC – включает наиболее часто используемые физические и химические данные, характеризующие органические соединения, и может выступать как обширный источник спектральной информации о соединениях (веществах). Физические данные - температуры кипения и плавления, плотность, растворимость, кристаллографические характеристики, цвет, удельный вес и т.п.; спектральные – ИК-, КР-, УФ-, ЯМР- и масс-спектры. Источник информации в этом случае – CRC Handbook of Data on Organic Compounds. Поиск в файле HODOC можно проводить по названию соединения, регистрационному номеру CAS, свойствам, терминам, физическим и спектральным данным.

SPECINFO. Эта база данных содержит информацию о спектрах ядерного магнитного резонанса (ЯМР) на ядрах углерода (^{13}C), фтора (^{19}F), фосфора (^{31}P), азота (^{15}N) и кислорода (^{17}O), а также инфракрасные (ИК) и масс-спектры соединений. Дополнительно к спектральным данным приводятся некоторые сведения, характеризующие соединение: молекулярная масса, элементный состав, название, регистрационный номер CAS. Особенность SPECINFO состоит в том, что в ее состав входит ряд специальных программ: CHESS – для поиска в БД соединений с идентичной или похожей на заданную структурой; COUPCAL – для расчета констант спин-спинового взаимодействия; GETSPEC – для поиска спектров соединения, соответствующего данной структурной формуле; SPECAL – для оценки параметров спектра ЯМР заданной структуры.

В аналитической практике в зависимости от типа решаемой задачи и характера

требуемой информации возможно использование не только перечисленных БД, но и ряда других, представленных в химическом кластере STN International. Полный перечень БД STN и их тематическую направленность можно найти в INTERNET:

www.cas.org;

www.fiz-karlsruhe.de (русскоязычное зеркало на www.permcnti.ru);

sibstn.nioch.nsc.ru (русскоязычная версия).

Для специалистов, занимающихся идентификацией органических соединений на основании спектральных данных, наибольший интерес представляют базы SPECINFO и HODOC. В качестве примера укажем, что по состоянию на 1.08.2001 SPECINFO содержит более 65 тыс. масс-спектров, ~17 тыс. ИК спектров, ~80 тыс. спектров ^{13}C-ЯМР. В целом в этой БД представлена спектральная информация о ~150 тыс. соединений. Технология работы с БД STN International и объем хранящихся в ее химическом кластере сведений таковы, что можно в большинстве случаев найти достаточно полные сведения по интересующему вопросу. Опытный или прошедший обучение поиску информации специалист способен быстро отыскать сведения, релевантные запросу исследователя (аналитика, спектроскописта, химика-органика). Желательно сделать это при минимальных затратах на оплату соответствующих услуг. Именно поэтому в нашей стране создается сеть Центров, специалисты которых обеспечивают квалифицированный поиск информации в телекоммуникационной сети STN International. Персональное обращение к БД STN International невыгодно, оно ведет к излишним затратам времени и дополнительным расходам на оплату услуг по поиску информации в соответствующих базах.

Характер основных запросов исследователей можно условно разделить на две группы. Первая – поиск фактографической информации о соединении, например, химический сдвиг атома в заданном окружении в спектре ЯМР, температура кипения или плавления некоторого соединения и т.д. Вторая – поиск библиографической информации, например, публикаций некоторого автора или первоисточников, отбираемых по ключевым словам или по структуре (по регистрационному номеру) соединения. Достоинство систем коллективного пользования состоит в том, что в одном сеансе связи можно, обращаясь к различным файлам, отыскать все имеющиеся сведения, относящиеся к обеим группам.

Наряду с охарактеризованными выше БД универсального характера, для исследователя, решающего идентификационные задачи, доступны и необходимы *БД персонального пользования*, связанные, как правило, лишь с одним из методов анализа. Их содержание более специфично. Например, единичная запись может включать название соединения, структурную формулу, эталонный спектр конкретного вида (или характеристику хроматографического удерживания) и соответствующий источник информации. Отсюда и

характер решаемых с помощью этих БД частных задач, например, поиск спектра заданного соединения или хроматографических параметров, характеризующих это соединение, а также опознание его по экспериментальным данным, например по спектру поглощения или индексу удерживания. Информация, содержащаяся в БД персонального пользования, может быть использована для решения идентификационных задач, если пользователь имеет в своем распоряжении специальную информационно-поисковую систему (см. раздел 1.2).

Рассмотрим более подробно некоторые специализированные базы данных, соответствующие важнейшим аналитическим методам, используемым при решении идентификации органических соединений.

1.1.2. Базы данных по масс-спектрометрии

Среди методов исследования органических веществ масс-спектрометрия занимает особое положение. По таким параметрам, как универсальность и чувствительность, масс-спектрометрия, по-видимому, превосходит другие аналитические методы. Масс-спектр может быть получен практически для любого соединения в любом агрегатном состоянии (газообразное, жидкое, аморфное или кристаллическое). Анализ масс-спектрометрических данных позволяет выявить информацию, бесценную для решения идентификационных задач, в частности, определить молекулярную массу, характерный тип распада молекулярных ионов, изотопный состав. При этом требуются ничтожные количества исследуемого вещества (пробы). Техника регистрации масс-спектров на современных приборах позволяет получать все необходимые данные за доли секунды. Это обстоятельство, наряду с высокой чувствительностью, способствовало созданию гибридных методов, прежде всего - хроматомасс-спектрометрии. Здесь эффективное разделение сложной смеси органических соединений (газовый или жидкостной хроматограф) сочетается с идентификацией компонентов смеси (комплекс масс-спектрометр-ЭВМ).

Среди различных способов ионизации молекул (лазерная и полевая десорбция, термическая, химическая ионизация и т.п.) наибольшее распространение в практике масс-спектрометрического анализа нашла ионизация пучком электронов (электронная ионизация). Подавляющее большинство исследований, а также опубликованных коллекций и атласов масс-спектров связано именно с этим способом ионизации. Наиболее известные и опубликованные в виде справочных изданий коллекции содержат десятки тысяч эталонных масс-спектров самых разнообразных органических соединений. Например, атлас [17] включает ~19 тыс. спектров, каталог EPA/NIH [18] с дополнениями [19] ~40 тыс. спектров. В издании [20] представлено около 58 тыс. полных масс-спектров. Число сокращенных спектров в издании [21] близко к 70000. Известны и другие, менее представительные

собрания спектров, например [22-26].

Перечисленные атласы и каталоги продолжают находить применение при анализе спектров “вручную” на основе традиционного для масс-спектрометрии приема - поиска спектральных аналогий. Однако специалистам хорошо известно, что положительный результат достигается таким способом лишь при поиске масс-спектра конкретного соединения, т.е., когда заранее высказана гипотеза о возможном строении изучаемого вещества. Накопленные масс-спектрометрические знания (спектроструктурные корреляционные зависимости) можно эффективно использовать лишь в тех случаях, когда известно, к какому классу химических соединений относится исследуемое вещество, а перед аналитиком стоит лишь частная задача уточнения структуры соединения.

Действительно, закономерности распада молекул в масс-спектрометрии чрезвычайно сложны, носят скрытый характер, могут резко меняться при сравнительно небольших различиях в структуре соединений. Химия возбужденных ионов и ионных частиц самого разнообразного строения, образующихся при диссоциативной ионизации в газовой фазе, строго говоря, не изучена. В отличие от других спектральных методов (например, ИК и ЯМР-спектроскопии), теоретические воззрения практически бессильны при попытке их использования для моделирования или объяснения масс-спектров. Сформулированные закономерности и правила, таблицы спектроструктурных корреляций пока что настолько несовершенны, что их применение при анализе спектра неизвестного соединения вынуждает выдвигать слишком много подлежащих последующей проверке гипотез о строении неизвестного. Сложность интерпретации накопленных экспериментальных данных и спектроструктурных корреляций, как и недостаточный уровень информационного обеспечения, надолго затормозили развитие масс-спектрометрической идентификации, несмотря на большие потенциальные возможности этого метода. При опознании индивидуальных соединений исследователи, как правило, ограничивались сведениями о молекулярной массе, брутто-формуле и простейшими данными, извлекаемыми из той части масс-спектра, которая описывает первичные стадии деструкции молекулярных ионов.

Характерная особенность масс-спектрометрии – чувствительность спектра к условиям его получения. Это усложняет использование метода в качественном анализе. Масс-спектральная информация практически бесполезна для решения идентификационных задач, если в соответствующем источнике не указаны условия регистрации эталонных спектров. Масс-спектры одного и того же соединения, будучи записаны в разных условиях, могут значительно отличаться друг от друга. На вид спектра, полученного при ионизации молекул электронным пучком, оказывают влияние конструктивные особенности прибора, тип используемой системы напуска, ее температура, тип и скорость развертки спектра и т.д. (см.

например [27-30]. Спектр может измениться из-за наложения пиков примесей, присутствующих в пробе (даже если проба дает единственный хроматографический пик) или образовавшихся в системе напуска и ионизационной камере. Спектры одной и той же пробы могут заметно различаться даже в случаях тождественности заданных условий регистрации, если такие спектры записаны в различное время. Заметим в этой связи, что иногда не менее похожими оказываются спектры различных соединений!

Искажения спектра, вызванные техническими факторами, например, типом развертки или анализатора, иногда удается учесть, но в общем случае вопрос тождественности спектров и, следовательно, соединений при традиционном "ручном" анализе остается открытым. Еще большие трудности возникают в тех случаях, когда структура изучаемого соединения неизвестна, а набор гипотез о его возможном строении велик.

Создание *компьютерных* баз масс-спектрометрических данных и информационно-поисковых систем на их основе позволило существенно облегчить сопоставление масс-спектров и повысить надежность идентификации органических соединений. Этому процессу способствовал дискретный характер и цифровая форма исходной масс-спектрометрической информации. Единичный масс-спектр включает массовые числа пиков ионов (m/z) и их относительные интенсивности (J), то есть десятки, а то и сотни поисковых признаков.

Масс-спектрометрические БД содержат десятки тысяч эталонных спектров. Например, БД полных масс-спектров EPA/NIH/NIST содержит свыше 62 тыс. тщательно проверенных спектров индивидуальных органических веществ, база данных НТЦ ХИ при Новосибирском институте органической химии СО РАН – около 52 тыс. спектров. Большая БД сформирована группой Мак-Лафферти (F.W. McLafferty). По данным [31] она содержит около 140 тыс. спектров для 120 тыс. соединений. Крупными коммерческими поставщиками предприняты усилия по объединению баз данных различных организаций с целью снабжения ими современного спектрального оборудования. Унифицированные базы данных насчитывают до 200 тыс. записей и при наличии финансовых средств могут быть поставлены (закуплены) как составная часть аналитического оборудования. В работе [32] приводятся сведения о приблизительно тридцати других базах данных, существенно меньших по числу записей. Они могут представлять не меньший интерес, чем "большие" БД, так как содержат сведения о спектрах соединений, важных с точки зрения конкретного пользователя (например, о спектрах особо опасных токсикантов). Известны также БД сокращенных масс-спектров, содержащие информацию не обо всех, а лишь о наиболее интенсивных или информативных пиках каждого спектра.

Современные информационные технологии позволяют быстро отобрать масс-спектры, наиболее похожие на заданный. При этом обычно предполагается, что спектр пробы

принадлежит индивидуальному соединению. Сопоставление его с эталонными спектрами индивидуальных соединений из базы данных производится по определенному алгоритму, зачастую неизвестному для пользователя, который вынужден доверять этому алгоритму и может проконтролировать лишь окончательный результат. Очевидно, что в этой ситуации качество спектральных данных играет чрезвычайно важную роль [33,34]. Этот вопрос рассмотрен в разделе 1.1.6.

1.1.3. Базы данных по ИК-спектроскопии

Сопоставление зарегистрированного пользователем ИК спектра поглощения пробы с опубликованными ранее эталонными ИК-спектрами стало одним из самых распространенных приемов установления состава и строения веществ. Информативность ИК спектра настолько высока, что при полном совпадении спектра индивидуального соединения с одним из эталонных спектров можно с немалой уверенностью утверждать, что соединение идентифицировано. Именно это обстоятельство и привело к созданию больших коллекций ИК спектральных данных. Интерес к этой деятельности в значительной степени “подогревается” доступностью метода, его высокой чувствительностью и особенно широтой использования в аналитической практике. Так, например, при обсуждении в Японии вопроса о создании баз данных по различным видам спектроскопии, Национальная химическая лаборатория промышленности пришла к выводу о необходимости первоочередного создания БД именно по ИК спектрам молекул [35].

За полувековой период использования ИК спектроскопии накоплен обширный материал о колебательных спектрах различных соединений. Сформулированы эмпирические правила, характеризующие корреляционные связи фрагментов молекул и соответствующего спектрального отклика. Разнообразные коллекции данных: атласы спектров, картотеки, каталоги на микрофишах и др. в сочетании с известными спектроструктурными корреляциями [36-39] и в “докомпьютерный” период оказывали помощь химикам и спектроскопистам в интерпретации ИК спектров. В известном руководстве по ИК спектроскопии [39] перечислено свыше 50 источников спектральных данных. В специализированной библиотеке спектральной информации Новосибирского института органической химии СО РАН собрано свыше 300 книг, атласов и каталогов колебательных спектров молекул. Некоторые коллекции ИК спектров, зарегистрированных на призменных и решеточных спектрометрах, например коллекция Садтлера [40], общепризнаны и в настоящее время могут считаться образцовыми. Эта коллекция и сегодня продолжает пополняться ИК спектрами соединений в конденсированной или в газовой фазе. Однако все более значимыми, емкими и распространенными становятся компьютерные БД (см. табл.

1.1.), содержащие, наряду с опубликованными ранее, спектры, зарегистрированные с гораздо большей точностью на современных приборах (преимущественно фурье-спектрометрах, FT-IR).

Публикации, посвященные использованию ЭВМ для создания баз данных по ИК-спектроскопии, а также математическим методам анализа этой информации для установления строения соединений, появились еще в конце 60-х годов. В этот период формируются два основных направления исследований. Первое – создание систем на основе “искусственного интеллекта”, получивших в последнее время название *экспертные системы*. Второе – создание систем на основе фактографических баз ИК спектральных данных.

Таблица 1.1.

Базы данных ИК спектров, содержащие информацию о структурных формулах

Коллекция	**Число спектров**	**Примечание**
Sadtler	160 000	часть FT-IR
Sadtler vapor phase	9 200	FT-IR
Canadian Scientific Numerical Database Service	166000	
НТЦ ХИ при НИОХ СО РАН	Свыше 70000	полные спектры, около 50000 структур
Aldrich-Nicolet	17 000	**FT-IR**
Sigma-Nicolet	10 600	- - -
Aldrich vapor phase	5 000	- - -
NIST/EPA vapor phase	5 244	- - -
NIMCR Japan	46 400	- - -
SpecInfo	22 600	17 000 полных спектров, 6600 положений полос
Coblentz Society	10 500	4 400 полных спектров

IRDC Japan	19 000	длины волн и интенсивности полос поглощения
Sprouse Scientific	Несколько небольших коллекций	- - -

Развитие первого направления рассмотрено в ряде монографий и обзоров [4, 7-9, 41-45], включая обстоятельный обзор достижений за последнее десятилетие, представленный М.Е. Эляшбергом [46]. Иначе обстоит дело с обзорами по второму направлению. Вероятно, наиболее детальное обозрение работ в этой области ИК спектроскопии сделано Луинжи [43] в 1990 году. Более поздние обобщения литературных данных, за редким исключением [47], рассматривают лишь вопросы совместного использования БД по различным видам спектроскопии (например [44,45]).

По способу доступа ИК спектральные БД можно с определенной долей условности разделить на централизованные и персональные. Централизованные – крупные БД, создатели которых стремятся включить в их состав как можно больше доступной информации. Как отмечалось ранее, такие БД создаются в коммерческих целях (поставляются вместе с современным спектральным оборудованием), а также для коллективного пользования; в последние годы часть из них доступна по сетям ИНТЕРНЕТ. Персональные БД создаются, пополняются и используются непосредственно их владельцами. Отбор спектров в эти БД производится в соответствии с научными интересами создателей [42, 48-50].

Объединение персональных БД или их присоединение к существующим централизованным БД не всегда возможно. Причина состоит в том, что в случае ИК спектроскопии локальные БД сильно различаются по качеству спектральной информации, оценка которого компьютерными средствами до настоящего времени затруднена. Компании, занимающиеся созданием коммерческих БД, и международные организации (CODATA, IUPAC), заинтересованные в развитии существующих централизованных БД, разработали и предлагают стандарты записи спектров ИК поглощения [51] и стандарты обмена спектрами [52-54], например, JCAMP-DX [52]. Эти рекомендации используются сегодня большинством фирм – производителей спектрометров.

Поддержка, пополнение существующих централизованных БД и расширение доступа к ним стоят чрезвычайно дорого [55]. Они не сулят научных лавров [56] и невозможны без государственных дотаций. Так, БД 5500 полных ИК спектров, созданная к 1985 г. при

помощи правительственных субсидий, не была пополнена до 15000 спектров (в соответствии с планом) из-за прекращения государственного финансирования [56]. Как указывается в [57] “экономические проблемы численных БД связаны с тем, что данных слишком мало и слишком много”. При небольшом объеме БД трудно привлечь исследователей к активному ее использованию, в то же время стоимость хранения, актуализации крупных баз и поиска в них информации возрастают с увеличением их объема. На практике наиболее доступны проблемно-ориентированные БД, содержащие, как правило, лишь сотни записей, относящихся к определенной тематике. Ограниченный размер доступных и качественных спектральных БД в этом случае препятствует прогрессу в области разработки автоматизированных методов для установления строения новых соединений. Число исследовательских групп, активно занимающихся этой тематикой, можно перечислить по пальцам. В то же время можно с уверенностью утверждать, что крупные БД с соответствующим программным инструментарием “будут широким средством анализа в 21 веке” [32], если они позволят решать набор задач, выходящий за рамки тривиального поиска спектра тождественного соединения.

Современные базы данных по ИК спектроскопии наряду с полной спектральной кривой содержат информацию о структурных формулах соответствующих индивидуальных соединений и ряд других сведений, сопровождающих соединение. Наличие в БД информации о структурной формуле соединения и его регистрационного номера позволяет не только решать задачу оперативного обслуживания специалистов необходимыми справочными данными, но и обеспечивает возможность проверки эффективности самих поисковых алгоритмов, новых теоретических описаний структур и изучения взаимосвязи свойств и строения молекул. Наконец, наличие кодов структурных формул необходимо, как будет показано в главе 2, для извлечения как можно более полной информации о строении исследуемых соединений (с использованием соответствующих алгоритмов и программного обеспечения).

При создании БД по ИК спектроскопии необходим комплекс программных и технических средств, позволяющий вводить и контролировать всю относящуюся к спектру и соединению информацию. Так, например, в используемой в НТЦ ХИ технологической схеме процесс создания БД на основе литературных данных делится на три потока:

- ввод структур соединений с помощью специализированного устройства планшетного типа “Граф” [49] или структурного редактора;
- ввод спектра с помощью комплекса “Спектр” (на начальных стадиях формирования базы). Позднее для этой цели использована технология сканирования, с программным обеспечением распознавания спектра;

- ввод дополнительной алфавитно-цифровой информации, характеризующей соединение и спектр.

Все записи, относящиеся к одному соединению, сопровождаются единым ключом (регистрационным номером), позволяющим “сливать” их при формировании БД. К началу 2001 года объем базы данных по ИК-спектроскопии в НТЦ ХИ составлял около 75 тыс. записей.

Основные концепции создания компьютерных БД (отбор, оцифровывание данных, контроль качества спектров, характер дополнительной информации и форматирование данных), рассмотренные в работе [47], не устарели до сих пор, но функцию оцифровывания спектра взял на себя современный спектрометр. Сведения об условиях регистрации, структуре соединения и другую сопроводительную информацию обычно хранят в отдельных файлах. Организация взаимодействия между файлами зависит от поставленных задач.

Созданию БД, содержащих полное описание спектральной кривой, уделяется сегодня все более пристальное внимание. Работы в этом направлении стимулируются не только наличием в широкой практике измерительно-вычислительных комплексов спектрометр-ЭВМ, доступностью средств ведения персональных баз данных, но и потребностью аналитической практики. Высокая стоимость крупных баз по ИК спектроскопии позволяет закупать их для персонального использования лишь в исключительных случаях. В аналитических лабораториях, как отмечено выше, используют ограниченные по числу записей, и поэтому более дешевые проблемно-ориентированные БД, например, спектры красителей, мономеров и полимеров, лекарственных препаратов, наркотических средств, сельскохозяйственных препаратов, основных загрязнителей окружающей среды и т.п. Такие БД насчитывают обычно от нескольких сотен до нескольких тысяч спектров веществ.

1.1.4. Базы данных по спектроскопии ядерного магнитного резонанса

Отличительная особенность спектроскопии ядерного магнитного резонанса (ЯМР) по сравнению с другими методами, например масс-спектрометрией и ИК спектроскопией, состоит в том, что в этом случае каждый сигнал спектра обусловлен резонансом соответствующего ядра (атома, спина) в его индивидуальном окружении. Мультиплетность сигналов, характеризующая константы спин-спинового взаимодействия, интенсивность и положение сигналов в шкале химических сдвигов дают, как правило, достаточно полное представление о ближайшем, а иногда и более далеком окружении резонирующих ядер. Полный спектр – совокупность всех сигналов – есть характеристика молекулы, а его индивидуальные компоненты – характеристика данного атома, типов его связей с соседними атомами и типов самих соседей. В этом состоит основное отличие спектров ЯМР от

рассмотренных ранее; поскольку ИК- и масс-спектры описывают поведение молекулы как единого целого. Но в этом и состоит огромное достоинство метода ЯМР – прозрачность и простота спектроструктурных корреляционных зависимостей.

Наиболее широко в химической практике используется спектроскопия ЯМР на ядрах водорода ^{1}H и углерода ^{13}C. В этих областях спектроскопии со времени активного внедрения метода в практику накоплен огромный экспериментальный материал. Казалось бы, простота восприятия и анализа данных ЯМР, ясные спектроструктурные зависимости, с одной стороны, и совершенствование экспериментальных методик, например, быстрое распространение двумерной ЯМР, с другой, - ставят под сомнение необходимость создания баз данных по этому виду спектроскопии. В пользу этого же свидетельствует зависимость вида спектра от рабочей частоты соответствующего инструментария. Например, спектры ^{1}H-ЯМР, записанные ранее на приборах с частотами 60-100 Мгц и хранящиеся на типографских носителях, имеют часто совершенно иной вид, чем зарегистрированные на современных приборах с частотами 300-700 Мгц. В меньшей степени подвержены этому влиянию спектры ^{13}C-ЯМР. Совершенствование инструментальной базы и расчетных методик, наряду с возможностью построения по данным ЯМР остова молекулы путем выявления последовательностей связей ^{13}C-^{13}C и взаимосвязей ^{13}C-^{1}H, ^{13}C-^{19}F, ^{13}C-^{31}P и т.п., создание и разработка эффективных экспертных систем, использующих не базы данных, а базы знаний (таблицы спектроструктурных корреляций), также противоречат необходимости создания соответствующих БД, что и высказывается рядом специалистов по ЯМР. Но дело обстоит не так просто.

Вероятно, здесь уместно еще раз подчеркнуть, для каких целей создаются базы данных. Действительно, опытный специалист в области ЯМР спектроскопии способен решить задачу установления строения изучаемого соединения без компьютера и базы данных. При этом он использует как собственный багаж знаний, так и прибегает к постановке специальных и дорогих экспериментов, применяя (в том числе) уникальное спектральное оборудование или расчетные методы. В то же время в массовой практике, как правило, анализируются ранее изученные объекты природного или антропогенного происхождения. Здесь используют существенно более дешевое оборудование, обладающее ограниченными возможностями, а соответствующий персонал часто не владеет необходимым объемом знаний и набором экспериментальных или расчетных средств. В этих условиях база данных и соответствующая ИПС становятся незаменимыми, простыми и общедоступными средствами идентификации. Вероятно, это и служит основной причиной неослабевающего интереса к созданию и наполнению новой спектральной информацией баз данных по спектроскопии ЯМР. Заметим дополнительно, что такие БД служат источником знаний,

используемых, в том числе, при совершенствовании экспертных систем. Основные источники информации, заносимой в БД: опубликованные коллекции спектров, руководства по ЯМР, периодические издания, локальные коллекции исследователей. Ряд наиболее крупных коллекций спектров ^{1}Н- и ^{13}С-ЯМР перечислен в табл. 1.2. Другие, достаточно представительные и специализированные подборки спектров опубликованы, например, в работах [65-67].

Исторически сформировалось два основных метода построения БД по ЯМР спектроскопии. Первый – ввод и хранение в памяти ЭВМ спектральных параметров (величины химических сдвигов, константы спин-спинового взаимодействия, времена релаксации), характеризующих атом(ы) или группы атомов в фиксированном окружении. Второй – хранение на машиночитаемых носителях кода структуры соединения (матрицы смежности) и параметров спектра, отнесенных к вершинам соответствующего химического графа (молекулярного графа).

Таблица 1.2.

Некоторые коллекции спектров ЯМР

№	Название коллекции	Количество спектров	Ссылка
1	Указатель литературных данных по спектроскопии ЯМР-^{1}Н	~60000 ^{1}Н (табличные данные)	[58]
2	The Sadtler Standard Spectra. Nuclear Magnetic Resonance Spectra. Vol. 1-94	52000 ^{1}H	[59]
3	Pouchert C.J., Behnke J. The Aldrich Library of 13C and 1H FT-Spectra. Vol. 1-3.	12000 ^{1}H, 12000 ^{13}C	[60]
4	Handbook of Proton-NMR Spectra and Data Asahi Research Center. Vol. 1-10	8000 ^{1}H	[61]
5	The Sadtler Standard Carbon-13 Nuclear Magnetic Resonance Spectra. Vol. 1-160	32000 ^{13}C	[62]
6	Bremser W., Hardt A., Ernst L. Carbon-13 NMR Spectral Data	~58000 ^{13}C	[63]

7	Grasselli J.G., Ritchey W.M. Atlas of Spectral Data and Physical Constants for Organic Compounds. Vol. 1-6	~20000 ^{1}H ~1000 ^{13}C	[64]

Ввод графической информации (структурные формулы, фрагменты) требует или разработки специальных технических средств (см., например [49,68]) или соответствующего программного обеспечения. Рассмотрим в качестве примера автоматизированное рабочее место (АРМ) со встроенным структурным редактором, используемое в НТЦ ХИ при создании БД по ЯМР спектроскопии на основе разнообразных литературных источников. АРМ позволяет вводить данные, проводить диагностику ошибок, просматривать, хранить, отыскивать требуемые сведения. Упрощенная функциональная схема приведена на рис. 1.1. АРМ снабжен средствами гибкой настройки окна системы для организации макета вводимых данных с учетом имеющейся в источнике информации. Ряд сервисных функций, позволяет маркировать вершины графа с целью однозначного отнесения спектральных параметров (химический сдвиг, интенсивность, время релаксации, константы спин-спинового взаимодействия) и автоматически заполнять повторяющиеся поля (типично для условий регистрации). Он снабжен справочниками по библиографии, растворителям, стандартам, спектрометрам и т.д. Все это и обеспечивает быстрое наполнение соответствующей БД.

(рис1.1 есть и в виде отдельного файла)

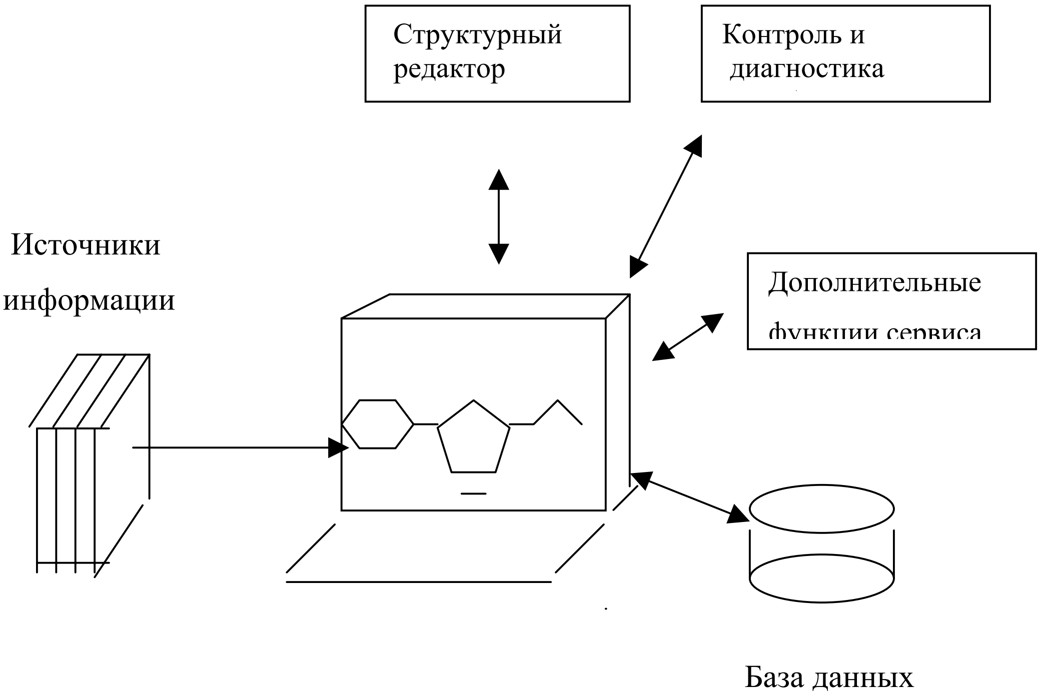


Рис. 1.1. **Функциональная схема рабочего места оператора при наполнении БД по спектроскопии ЯМР**

Наиболее полная форма БД по спектроскопии ЯМР может содержать и спектральную кривую в оцифрованном виде, что иногда и делают в случае ^{1}Н-ЯМР. В случае ^{13}С-ЯМР спектр, как правило, представляют в дискретной форме (химический сдвиг, интенсивность сигнала). Сведения о полном оцифрованном спектре сегодня доступны большинству спектроскопистов, использующих современную технику ЯМР. Однако в силу ряда причин, часть из которых обсуждена выше, небольшие БД этого типа стали создаваться лишь в последнее время. Ниже кратко охарактеризованы наиболее известные БД по спектроскопии ядерного магнитного резонанса.

База данных FIZ K. Это, вероятно, наиболее широко известная и используемая база. Доступ к ней (см. раздел 1.2) обеспечивается через международную линию связи STN International. В России квалифицированный и экономичный поиск информации в этой базе возможен, в частности, через специализированные региональные отраслевые или академические Центры. По состоянию на конец 2000 года БД FIZ K содержит 152 тысячи химических соединений и 90 тысяч спектров различных типов. Основную часть составляют спектры ^{13}С-ЯМР. Информация собирается из более 20 основных журналов химического профиля, таких как “Organic Magnetic Resonance”, “Journal of American Chemical Society” и т.п., ряда спектральных каталогов и коллекций, например, фирм BASF, Bruker и других. В этой БД содержатся сведения о названии соединения, его регистрационном номере по CAS, структурной и молекулярной формуле, спектральные характеристики сигналов (химические сдвиги, константы, интенсивности, мультиплетности), условия записи спектра и библиографические сведения. База данных и поисковое программное обеспечение поддерживаются системой SpecInfo и предоставляют широкие возможности пользователю (см. раздел 1.1.1).

БД исследовательской лаборатории Садтлера. База данных спектров этой лаборатории хорошо известна и включает 52000 спектров протонного магнитного резонанса и более 60000 спектров ^{13}С-ЯМР. По данным [69] база ^{13}С-ЯМР содержит информацию о химически чистых соединениях, включая оцифрованный спектр. Программное обеспечение реализовано на IBM PC. Коллекция спектров ^{1}Н-ЯМР этой лаборатории, описанная в справочнике [70], используется в файле HODOC, доступном через INTERNET (см. 1.1.1).

БД Научно-технического центра химической информатики. Базы данных по ^{1}Н и ^{13}С ЯМР спектроскопии сформированы на основе информации, взятой из опубликованных каталогов спектров и научных журналов химического профиля, издаваемых на русском языке. Для каждого соединения хранится следующая информация: топологический код структурной формулы; химические сдвиги сигналов с отнесением к элементам структуры;

мультиплетность и интенсивность сигналов; регистрационный номер соединения; название соединения; молекулярный вес и молекулярная (брутто) формула; координаты узлов структурной формулы для ее графического воспроизведения на дисплее или печати.

БД ^{1}H-ЯМР содержит в этом случае спектральные сведения о 48000 соединений из коллекции [59] и ~20000 из коллекции НТЦ ХИ [58]. База ^{13}C-ЯМР включает 27 тыс. записей из коллекции [62]. По экономическим причинам эти БД в последние годы не пополнялись.

Другие базы данных. Наряду с перечисленными, в литературе приводятся сведения о других достаточно крупных БД. В работе [71] сообщается о БД химической информационной системе CIS, содержащей файл спектров ^{13}C-ЯМР объемом 11700 записей. Как следует из [72] пополнение базы прекращено. База данных Японской национальной химической лаборатории содержит по данным [72,73] 6000 ^{1}H и ^{13}C ЯМР спектров, записанных в тщательно контролируемых условиях. Шанхайский институт органической химии создал БД, содержащую более 20 тыс. спектров ^{13}C-ЯМР. Структурная формула соединений в этом случае представлена в коде Висвессера. В работах [74,75] приведена информация о БД Аризонского университета и университета Вены. Последняя, по данным [32], насчитывает уже около 90000 ^{13}C спектров. В системе SPIRES [76] используется БД, включающая ^{1}H- и ^{13}C-ЯМР спектры свыше 10 тыс. органических соединений. Покупателям приборов фирмы Брукер доступна БД на компьютерах Aspect. Сведения о ряде других небольших БД можно найти в обзоре [32]. В последнее время все более пристальное внимание уделяется вопросам создания баз данных на других ядрах [68].

Обзор коллекций и баз данных показывает, что, несмотря на широкую распространенность ^{1}H-ЯМР в аналитической практике, основное внимание исследователей и поставщиков БД сконцентрировано на создании БД по спектроскопии углеродного магнитного резонанса. Это обусловлено рядом причин: в первую очередь – высокой селективностью ^{13}C-ЯМР спектров. В диапазоне химических сдвигов более 200 м.д. каждому пику спектра в условиях полного подавления спинового взаимодействия с протонами соответствуют индивидуальные атомы или группы магнитно эквивалентных атомов молекулы. Положения пиков (обычно приводятся с точностью до двух знаков после запятой) хорошо коррелируют с химическим окружением атомов. Высокой информативностью обладают константы спин-спинового взаимодействия. Поэтому сознательно отбираемая при поиске информация (из совокупности данных, сконцентрированных в БД) способствует быстрому анализу новых экспериментальных данных. Не случайно, для решения справочных задач часто используют не полные фактографические базы, а сформированные на их основе компьютерными средствами таблицы спектроструктурных корреляционных зависимостей. Последние представляют собой либо иерархическое описание спектральных параметров

(например, химические сдвиги, отвечающие тем или иным атомам в соответствующем окружении), либо, наоборот, упорядоченное описание кодов фрагментов с соответствующими центральному атому фрагмента параметрами спектра.

Несмотря на накопленный огромный экспериментальный материал, число созданных и используемых компьютерных БД по спектроскопии протонного магнитного резонанса весьма ограничено. Лишь в самом конце девяностых годов описана “первая в мире цифровая ^{1}H-ЯМР база данных с полным спектральным образом” [77, 78]. Диапазон химических сдвигов составляет в этом случае около 10 м.д. и относительная полуширина сигналов становится сопоставимой с ним. Вид спектра связан с рабочей частотой спектрометра, которая претерпела существенные изменения в связи с совершенствованием инструментального парка и переходом на использование сверхпроводящих магнитов. До сих пор основной массив опубликованных ^{1}H-ЯМР спектров зарегистрирован на спектрометрах с рабочими частотами 60, 90 МГц; в повседневной же практике настоящего времени используют приборы с частотами 200 – 400 МГц и выше. Очевидно, простое сопоставление полного спектра пробы с целью поиска в БД формально подобных спектров можно проводить и в этом случае, но его результат требует осмысленного анализа. Вероятно, БД полных спектров ^{1}H-ЯМР могут эффективно использоваться для идентификации соединений лишь только тогда, когда механизм их применения позволяет пересчитывать вид спектра при переходе с одной частоты регистрации на другую, например, с более высокой частоты на меньшую. Принципиальная возможность этого имеется, поскольку на приборах с частотами 200 МГц и выше сложные спиновые системы часто трансформируются в системы первого порядка. Это позволяет легко определять точные значения химических сдвигов и констант спин-спинового взаимодействия и затем моделировать требуемый спектр, учитывая различия в частотах регистрации, с помощью хорошо известных подходов и программ. Переоценивать эту возможность, однако, не следует, поскольку, с одной стороны, трудно учесть все факторы, привносящие возможные искажения вида спектра ^{1}H-ЯМР, обусловленные изменениями химических сдвигов (растворитель, концентрация, температура и т.д.) особенно в случаях сильно связанных многоспиновых систем. С другой, не следует забывать, что, как правило, БД используют для идентификации соединений в массовой аналитической практике специалисты, которые могут и не владеть расчетным инструментарием. Поэтому в случае ^{1}H-ЯМР спектроскопии в ближайшие годы, вероятно, будут широко использоваться не БД полных спектров, а версии БД вида “молекулярный граф – отнесения спектральных признаков к вершинам графа”.

1.1. 5. Другие базы спектральных данных

Кроме вышеперечисленных, для решения идентификационных задач существуют и используются базы данных по другим спектральным методам: например, по атомно-эмиссионной спектроскопии в видимой и УФ области [79], по люминесцентной спектроскопии молекул [80], по ЭПР-, КР- и оже-спектроскопии. Объем соответствующих БД существенно меньше ввиду ограниченного перечня возможных компонентов. Так, лишь немногие органические вещества дают спектры флуоресценции. Наиболее представительный атлас Садтлера содержит 2000 спектров возбуждения и испускания флуоресценции. Структурные формулы эталонов в соответствующие БД часто не вносят, ограничиваясь названиями.

Спектры флуоресценции молекул при комнатной температуре малоинформативны и требуют специальной обработки при вводе в БД (корректировка, сглаживание, перевод в цифровую форму). Как и в ИК-спектроскопии, весь спектральный диапазон можно разбить на интервалы постоянной ширины $\Delta\lambda$ и хранить в БД среднее значение относительной интенсивности по каждому интервалу [81]; от выбора $\Delta\lambda$ и, следовательно, от числа каналов передачи информации зависит точность передачи спектра и достоверность опознания соответствующего вещества. На ранних стадиях развития компьютерных БД и соответствующих ИПС лимитирующим фактором для выбора $\Delta\lambda$ при передаче и опознании спектров являлся объем требуемой памяти, но с развитием компьютерной техники данный фактор стал менее значимым. Тем не менее "лобовой" подход при переводе сплошной спектральной кривой в дискретную цифровую форму и сейчас не является единственно возможным и не всегда оптимален для организации поиска.

Так, при записи в БД широкополосных спектров флуоресценции можно провести предварительное дифференцирование, отобрать точки, где производная $\partial I / \partial\lambda$ меняет знак или становится равной нулю, а затем ввести в БД значения $I_{отн.}$ и λ для этих точек. Другой возможный путь - разложение спектра флуоресценции в ряд Фурье. Для надежного опознания индивидуальных соединений (например, алкилфенолов) оказалось достаточно хранить в памяти ЭВМ информацию по 9 первым гармоникам каждого эталонного спектра [82]. Можно также разложить каждый эталонный спектр на гауссовы составляющие либо рассчитать четыре его центральных момента. Так же поступают со спектром исследуемой пробы. Значения параметров гауссиан или центральные моменты для спектров пробы и эталонов сопоставляются в ходе поиска. По мнению авторов [83] такие подходы дают более надежные результаты поиска чем обычная запись спектров или использование разложения в ряд Фурье. Отметим, что при небольшом объеме БД вещества хорошо опознаются даже при бинарной кодировке эталонных спектров, когда каждому значению λ ставится в

соответствие 0 (нет пика) или 1 (есть пик) [10]. Бинарная кодировка спектров неоднократно применялась и в атомно-эмиссионной спектроскопии [79], и в масс-спектрометрии. Этот способ записи спектра весьма устойчив к случайным погрешностям регистрации, присутствию примесей и т.п. Очевидно, проблема выбора оптимального способа кодировки эталонных спектров возникает при формировании любой БД и должна решаться так, чтобы обеспечить наиболее правильные и устойчивые решения идентификационных задач.

1.1.6. О качестве спектральных данных

Надежность опознания некоторого соединения по его спектру зависит не только от того, содержится ли эталонный спектр того же соединения в соответствующей БД. На результат поиска оказывают сильное влияние способ представления исходных данных, алгоритм поиска и способ сопоставления спектров. Основные же факторы - это *воспроизводимость и селективность спектральных данных*, которые, в свою очередь, зависят от целого ряда факторов, в частности, от типа спектра, качества работы спектральной аппаратуры и природы исследуемых веществ. Общепринято, что библиотека эталонных спектров должна содержать записи, относящиеся лишь к индивидуальным соединениям (чистым веществам). Их чистота должна быть подтверждена, например, методом хроматографии и/или другими данными (точка кипения, плавления и т.п.).

Идентификация может быть успешной только в тех случаях, когда спектр пробы и эталонные спектры соответствующего вида записаны в близких или тождественных экспериментальных условиях. Но даже в том случае, когда спектры одного и того же вещества записываются повторно на одной и той же аппаратуре и по той же методике, возможны небольшие различия в положении и относительной интенсивности полос (пиков). Наиболее низкая воспроизводимость характерна для масс-спектрометрии, наиболее высокая – для спектроскопии ЯМР. Для оценки воспроизводимости информации, содержащейся в БД, используются традиционные методы математической статистики, но они не в состоянии предсказать, насколько велики будут расхождения между эталонными спектрами одного соединения, записанными на разной аппаратуре или в несколько различающихся условиях; этот вопрос требует специальных межлабораторных исследований.

С другой стороны, в масс-спектрометрии или ИК спектроскопии достаточно часто возникает ситуация, когда спектры соединений различных химических классов, а тем более вещества одного и того же гомологического ряда, оптические и пространственные изомеры обладают очень похожими спектрами и, следовательно, не различаются даже при использовании самых изощренных поисковых алгоритмов. Разумеется, наблюдается и совершенно противоположная ситуация. К сожалению, селективность эталонных спектров

редко оказывается предметом специальных исследований.

Рассмотрим основные способы оценки качества спектральной информации, содержащейся в различных БД.

Масс-спектрометрия. Относительно низкая воспроизводимость масс-спектров обуславливает особые требования, предъявляемые как к данным поискового запроса, так и к процедуре регистрации эталонных спектров. Заключение о качестве экспериментального масс-спектра во многих случаях может сделать исследователь на основе известных эмпирических правил, с учетом следующих простых соображений. В спектре индивидуального соединения не должно быть пиков ионов с величинами *m/z*, превышающими *m/z* молекулярного иона (M^+) и его изотопосодержащих аналогов. Не должны проявляться нелогичные нейтральные потери, равные, например, 5 – 12, 21 – 25 а.е.м. и т.п. В случае хроматомасс-спектрометрии или хроматографии с масс-спектрометрическим детектором спектр должен быть записан на том участке индивидуального хроматографического пика, с такой скоростью развертки и регистрации, которые приводят к минимальным искажениям.

Качеству занесенных в масс-спектрометрические БД спектров пользователю приходится доверять. Производители БД, учитывая низкую воспроизводимость спектров этого вида, часто идут по пути включения в базу дублирующих друг друга спектров одного и того же соединения без их специального отбора. Критерием включения спектра в БД является его отличие от ранее занесенных. Поэтому в ряде баз данных встречаются спектры одних и тех же соединений, полученные из различных источников или записанные в различных экспериментальных условиях. В первую очередь это относится к наиболее подробно исследованным или часто встречающимся в аналитической практике веществам.

В ряде случаев для количественной оценки качества масс-спектров, заносимых или уже хранящихся в БД, используют специализированное программное обеспечение. Совокупность основных требований для расчета индекса качества Q , сформулированная в работах [76,84], опирается на опыт экспертов и, вероятно, будет использоваться еще достаточно долго. Контроль качества масс-спектра в этом случае осуществляется с помощью следующих параметров: источник спектральных данных, условия ионизации, наличие пиков примесей выше M^+ и его более тяжелых изотопных аналогов, нелогичные нейтральные потери, точность измерения изотопных пиков, общее число пиков в спектре, ограничение спектра по нижней области масс. Индекс качества определяется как произведение индивидуальных факторов качества *j*-признака (Q_j):

$$Q = \Pi\, Q_j$$

Здесь символ П - произведение величин Q_j. Последние рассчитывают по

эмпирическим выражениям так, что они изменяются в интервале от 0 до 1. При расчете индекса качества можно дополнительно учитывать чистоту образца, дату калибровки прибора, данные о калибровке спектра и т.п. Предполагается, что если результирующее значение параметра Q меньше 0.5, то такой спектр нельзя включать в БД, так как это может ухудшить результаты идентификации не только данного соединения, но и других соединений, имеющих спектры, близкие к данному.

Кратко рассмотрим наиболее важные составляющие величины Q_j. Параметру Q_1 – источник информации - присваивается значение, равное 1, если данные взяты из каталогов [17-19] или подтверждены как минимум двумя ссылками на независимые источники. Масс-спектры, опубликованные в других изданиях, характеризуют величиной Q_1, равной 0.95, неопубликованные - $Q_1 = 0.9$. В книге [30] такая оценка поставлена под сомнение. По мнению ее авторов, параметр Q_1 должен определяться не столько "доверием к источнику, сколько воспроизводимостью спектра в конкретных условиях регистрации". С этим нельзя не согласиться.

Воспроизводимость спектра (V) рассчитывают согласно выражению $V = \Sigma I^*\Delta I$, где I^* - средние значения интенсивностей в процентах к суммарному ионному току, ΔI – их доверительные интервалы в серии экспериментов. А величина $Q_1 = 1 - V / 1000$. В этом случае при $V \leq 100$, $Q_1 \geq 0.9$, а при $V \geq 500$, $Q_1 \leq 0.5$. Параметр Q_2 – условия ионизации - чрезвычайно важен. В первую очередь потому, что характер масс-спектра сильно зависит от энергии ионизирующих электронов. При энергиях 50-70 эВ масс-спектры наиболее интенсивны и стабильны. В этих случаях принимают параметр $Q_2 = 1$. При энергиях (E) менее 50 эВ параметр принимает значение, равное $Q_2 = (E-10) / 40$, где $10 \leq E \leq 50$ эВ.

Фактор Q_3 , учитывающий примеси веществ с молекулярной массой, превышающей массу основной компоненты, определяют как $Q_3 = [(I_M+5) - 2\Sigma I_x] / (I_M+5)$, где I_M – относительная интенсивность пиков молекулярных ионов, ΣI_x - относительная интенсивность всех пиков с массой, большей молекулярных ионов основной компоненты.

Условия ионизации необходимо учитывать и в тех случаях, когда в одной и той же БД хранят спектры, полученные, например, химической и электронной ионизацией, полевой десорбцией и т.п. Спектры одного и того же вещества в этих случаях драматически отличаются друг от друга. Остальные сомножители менее значимы, однако их также учитывают при оценке качества эталонного спектра, сопровождая индексом качества каждую запись. Использование индексов качества препятствует дублированию записей, позволяет заменять устаревшие данные новыми, более надежными, сравнивать качество используемых масс-спектров и ориентироваться при принятии окончательного решения по результату поиска.

Заметим, что при оценке качества записей, заносимых в базы данных по масс-спектрометрии, следует прибегать к помощи специалистов в этой области. Так делается, например, при формировании БД Национального бюро стандартов и технологии США.

ИК спектроскопия. Несколько хуже обстоит дело с оценкой качества ИК спектров. Известно, что в данном случае вид спектра сильно зависит от условий регистрации образца, типа прибора, фазы анализируемого образца: твердая форма, раствор или газ. Вероятно, можно уверенно говорить о высоком качестве записей БД, сформированной на основе данных лаборатории Садтлера. Но даже и в этом случае сложная технологическая цепь формирования БД, содержащей графическую и алфавитно-цифровую информацию о спектрах десятков тысяч соединений, и необходимость привлечения труда многих технических работников сопровождается появлением ошибочных данных как в опубликованных материалах, так и, по-видимому, в соответствующих БД. Лишь в самое последнее время предложены приемы оценки качества записей в БД, основанные на подобии спектров соединений, обладающих похожими структурами [77, 85]. Они пока несовершенны, не снабжают записи количественной мерой качества спектров, но, несомненно, будут развиваться.

Наиболее качественными сегодня можно считать ИК спектры в газовой фазе, полученные, в том числе на хромато-ИК-спектрометрах, для индивидуальных веществ. Во всех остальных случаях качество ИК спектров оценивает или эксперт (при формировании БД) или экспериментатор – пользователь ИПС. Роль эксперта (спектроскописта, аналитика) при оценке результатов использования БД в этом случае особенно важна. Визуальный анализ спектров способен часто помочь там, где не может удовлетворительно справиться поисковое матобеспечение.

Спектроскопия ЯМР. Мощь и привлекательность метода ЯМР в структурном анализе органических соединений такова, что химическая периодическая литература содержит огромный объем данных этого вида. К сожалению, как правило, они не полностью описывают спектры или отнесенные к ним фрагменты структуры соединения. Поэтому при формировании БД по ЯМР спектроскопии, содержащей качественную информацию, необходимо ответить на следующие вопросы. Целесообразно ли аккумулировать в БД сведения, если в источнике информации (например, журнале) для соединения приводится не полный спектр и соответствующий ему набор параметров, а лишь малая часть данных (например, один или два химических сдвига), обсуждаемые в статье? Следует ли использовать эти сведения, если в источнике не аргументируется отнесение сигналов к фрагментам структуры? Это может быть обусловлено как позицией редакции, так и позицией авторов, акцентирующих внимание на других вопросах.

Наиболее привлекателен подход повторной регистрации спектров соединений с однозначно установленной структурой. При этом должны быть соблюдены рекомендованные международными организациями условия регистрации и проведен тщательный анализ отнесения сигналов к фрагментам структуры соединения (см., например [77, 86]) в том числе с привлечением самых современных экспериментальных приемов, например, регистрации двумерных спектров ЯМР. Но этот путь чрезвычайно дорог, часто не реализуем, не позволяет суммировать накопленные ранее исследователями сведения. Несомненное достоинство, однако, заключается в том, что в этом случае база содержит абсолютно достоверную информацию, но – увы – ограниченное число записей.

В созданных ранее крупных справочных базах используют иные оценки качества записей, например, оцениваемый с помощью компьютера разброс значений химических сдвигов атома (или групп атомов) в аналогичном окружении. На рис.1.2 приведен пример молекулярного графа одного соединения и гистограмм распределения химических сдвигов ^{13}C-ЯМР (шкала по горизонтали) атомов углерода в аналогичном окружении. Проставленные слева и справа от гистограмм номера соответствуют нумерации атомов углерода в приведенном молекулярном графе. Цифры, отмечающие пиковое значение для каждой из гистограмм, указывают число примеров, по которым построена эта гистограмма, то есть число эталонных спектров ЯМР, относящихся к соединениям со сходной структурой (с тем же ближайшим окружением данного атома углерода). Все примеры взяты из БД по ^{13}C-ЯМР, сформированной в НТЦ ХИ.

Судя по приведенным гистограммам, разброс значений химических сдвигов в подобных случаях сравнительно невелик. Поэтому новые записи, содержащие отнесения одного из данных атомов в тождественном структурном окружении, но имеющие значения химических сдвигов, далеко выходящие за рамки подобных гистограмм, должны подвергаться сомнению, и все данные таких записей должны быть тщательно проверены.

Быстрое развитие экспериментальных возможностей ЯМР, в первую очередь методов селективной и двумерной спектроскопии, способствует повышению качества новых экспериментальных данных. В случае ЯМР, как правило, не требуется сохранения в БД дублирующей спектральной информации, если она зарегистрирована в тождествен ных условиях. Новые записи, полученные на более совершенных приборах, заменяют устаревшие. Однако в БД в качестве сопровождающей спектр информации необходимо указывать не только дату регистрации спектра и тип прибора. На вид спектра и спектральные параметры влияют растворитель, примеси, стандарт, температура, динамические процессы и

т. п. Вся эта алфавитно-цифровая информация должна сопровождать спектр, отбираемый при поиске, то есть воспроизводиться в поисковом машинном ответе и учитываться аналитиком при решении поисковых задач.

1.2. Информационно-поисковые системы для опознания соединения по его спектру

1.2.1. Поисковые ситуации и предварительный отбор эталонных спектров

Основное назначение ИПС по молекулярной спектроскопии – отобрать из соответствующей БД эталонные спектры, подобные спектру пробы. Методология такого поиска неоднократно обсуждалась в литературе для общего случая [7, 10] и на примере масс-спектрометрии [87-90]. При этом анализировались типичные поисковые ситуации:

- эталонный спектр опознаваемого соединения присутствует в БД, условия его регистрации близки к условиям регистрации спектра исследуемой пробы;
- эталонный спектр соединения присутствует в БД, однако получен при существенно иных условиях регистрации;
- эталонный спектр соединения имеется в БД, но он недостаточного качества, например, записан с ошибками или в присутствии примесей;
- анализируемая проба (неизвестное вещество) представляет собой смесь соединений, причем эталонные спектры индивидуальных соединений (всех компонентов смеси) содержатся в БД;
- анализируемая проба представляет собой индивидуальное соединение, но его эталонный спектр отсутствует в БД. Зато там имеются похожие спектры гомологов, изомеров и других структурных аналогов опознаваемого соединения;
- спектры структурных аналогов опознаваемого соединения или всех компонентов исследуемой смеси в БД отсутствуют.

Сложность ситуаций в приведенном ряду нарастает. Первые три ситуации могут привести к достоверной идентификации исследуемого соединения; четвертая - к идентификации всех или некоторых компонентов пробы (см. главы 3 и 4); пятая - к установлению особенностей строения исследуемого соединения (см. главу 2). Это удается потому, что обычно эталонные спектры структурнородственных соединений сходны друг с другом; справедливо и обратное заключение. Что же касается шестой задачи, то шансов на ее успешное решение довольно мало.

К сожалению, до обращения к ИПС аналитик зачастую не знает, имеется ли в БД

эталонный спектр опознаваемого соединения (или всех компонентов сложной пробы); в тех ли условиях записаны эталонные спектры, что и спектр пробы; а иногда не знает даже, является ли исследуемая проба индивидуальным соединением или смесью сложного состава. Поэтому в идеале алгоритм информационного поиска должен быть как можно более универсальным и давать правильные результаты независимо от характера пробы. Однако достичь этого исключительно трудно.

В разделе 1.1. были рассмотрены основные приемы представления сопоставляемых спектров и сокращения спектральной информации, в данном разделе мы остановимся на общей методологии поиска и методах отбора из БД записей, релевантных запросу. Алгоритмы сопоставления спектров особенно бурно развивались в 80-е годы. К этому же периоду относятся основные обзоры [7, 43, 87, 91-93], посвященные обсуждаемому вопросу. Остановимся на наиболее характерных чертах и принципах работы спектральных ИПС, разработанных для решении ситуационных задач №№ 1-3 из вышеприведенного перечня.

Независимо от того, с какими спектрами работает такая ИПС, можно выделить следующие стадии поиска:

- ввод запроса (спектра пробы и априорной информации о ней) и преобразование информации;
- предварительный отбор эталонных спектров.
- сопоставление спектра пробы со всеми отобранными эталонными спектрами в рамках алгоритма прямого, обратного или комбинированного поиска. При этом степень сходства оценивается количественно. Эта стадия поиска детально рассматривается в разделе 1.2.2,
- формирование ответа на запрос. Поисковый ответ спектральных ИПС обычно включает ранжированный перечень эталонных спектров, наиболее похожих на спектр пробы. Перечень ограничивается с помощью априорно заданных критериев (см. раздел 1.2.3). Для каждого отобранного эталонного спектра указываются названия и/или структурные формулы соединения. Обычно тут же приводятся значения параметра, характеризующего степень сходства спектров, источник информации об эталонном спектре, брутто-формула и другие сведения о соединении и его спектре.

Таким образом, можно составить обобщенную схему поиска:

спектр пробы $\Rightarrow$ запрос $\Rightarrow$ преобразование спектра $\Rightarrow$ выбор условий предварительного отбора $\Rightarrow$ отбор 1 $\Rightarrow$ окончательное сравнение спектров $\Rightarrow$ отбор 2 $\Rightarrow$ ответ

Достоверность и однозначность ответа определяются не только объемом и качеством информации, содержащейся в БД; результат поиска зависит от того, какая информация предъявлена в поисковом запросе. Типичный пример: если в БД содержатся сокращенные

эталонные ИК или масс-спектры, то и к спектру пробы должна быть предварительно применена адекватная процедура сокращения. В противном случае положительный результат поиска маловероятен.

Спектр пробы, как правило, представлен в оцифрованном виде в соответствующем выходном формате прибора. Эти форматы на современном оборудовании унифицированы, что позволяет, используя необходимые конверторы (интерфейс), передавать спектр в ИПС для формирования поискового запроса. Если же спектр пробы представлен в аналоговой форме или на бумажном (типографском) носителе, то запрос формирует сам пользователь, предварительно преобразуя данные в форму дискретного описания, например, составляя список пар чисел, характеризующих спектр: положение пика (сигнала) – интенсивность [94]. Затем эта информация вводится в ЭВМ с клавиатуры или иным способом.

Преобразование спектра пробы – процедура, выполняемая лишь для тех видов спектроскопии, в которых данные сильно зависят от условий регистрации. Она характерна для всех наиболее распространенных в практике масс-спектрометрических ИПС и поисковых систем по ИК спектроскопии молекул. Как отмечалось выше, его цель свести к минимуму влияние различий экспериментальных условий на результат поиска, отбора и фильтрации записей при минимальных затратах времени. При этом используют следующие простые приемы. Например, из полного масс-спектра удаляют малоинтенсивные пики или каждый из m спектральных интервалов описывают ограниченным числом (n) наиболее интенсивных пиков и др. Возможны различные сочетания параметров m и n [29,95]. Переменные значения параметров n и m используют и в системе Компас-МС [96], в этом случае n=3 в интервале малых величин m/z, n=2 – средних и n=1 для больших массовых чисел. Это сделано с целью выравнивания информативности пиков в различных областях спектра. Подобную же цель преследует и другой вид преобразования масс-спектрометрических данных – представление пиков в виде $m_iI_i^{1/2}$ [97]. В данном случае произведение m_iI_i отражает большую значимость пиков “тяжелых” ионов спектра, а степенная функция выравнивает влияние различий в интенсивностях. Интенсивности пиков в масс-спектрах или ИК спектрах достаточно часто представляют в кодированной (огрубленной) форме – ограниченным набором целых чисел (однобитовое представление 1-есть пик, 0 - нет; двухбитовое – 00 - нет, 01- слабый, 10 - средний, 11-сильный пик или полоса поглощения и т.д.). Эту же форму представления данных используют для предварительного отбора записей из БД. Иногда для целей предварительного отбора применяют такие преобразования спектра, которые полностью изменяют его вид по отношению к оригиналу. Например, полный спектр разбивают на несколько диапазонов, в каждом из которых проводят локальную нормализацию интенсивностей пиков. В работе [98]

спектр пробы представляют набором спектров, каждый из которых рассчитывают с учетом возможных различий спектров, полученных на спектрометрах, разных по методам разделения и регистрации ионов.

Разработчики ИПС стремятся найти компромисс между скоростью решения поисковой задачи и полнотой отбора из БД релевантных запросу сведений, т.е. совместить в одном программном продукте на первый взгляд несовместимые требования. Действительно, сопоставление полных спектральных записей в ряде случаев требует значительных временных затрат и иногда не может быть реализовано в режиме "on-line" (например, при отнесении хроматографических пиков по мере выхода соответствующих компонентов из колонки в спектральный детектор). Детальное сопоставление спектра пробы со всеми эталонными спектрами БД, в том числе со спектрами веществ, заведомо отсутствующих в пробе, не только бессмысленно, но и может приводить к недостоверным (или даже абсурдным) результатам поиска. Последнее возможно, если чистота исследуемого образца не гарантирована.

Практически любой алгоритм идентификации включает в себя стадию быстрого предварительного просмотра всех эталонных спектров, содержащихся в БД, с применением некоторых задаваемых разработчиком ИПС или пользователем критериев (дескрипторов). В результате предварительного отбора выделяется та часть спектральной БД, в которой наиболее вероятно присутствие опознаваемого соединения. В некоторый промежуточный файл (сокращенную базу данных, рабочую библиотеку) попадает ограниченное число "наиболее подозреваемых" эталонов. Естественно, условия предварительного отбора должны быть и жесткими и мягкими одновременно. Жесткими - для того, чтобы отобрать в список спектров, подлежащих детальному сопоставлению, минимальное число записей из используемой БД. Мягкими – чтобы не исключить из этого списка искомый спектр или спектры искомых соединений.

Один из способов предварительного отбора - отбраковка соединений, заведомо отсутствующих в данной пробе, на основании априорной информации. Этот прием используется практически во всех ИПС (не только спектральных), и он рассматривается как средство ускорения поиска. Вместе с тем такая отбраковка является способом повышения характеристичности поисковых признаков для тех эталонных спектров, которые остаются в сокращенной рабочей библиотеке [10]. Поэтому корректно проведенная предварительная отбраковка всегда приводит к более правильным и однозначным результатам качественного анализа. Для эффективной отбраковки в БД должна быть как можно полнее отражена информация по составу и структуре эталонов, их отнесению к тому или иному классу, условиям регистрации спектра и т.п. Соответствующая априорная информация о пробе

вводится пользователем одновременно со спектром этой пробы. Для этой же цели можно использовать сведения о молекулярной массе и брутто-формуле соединения. В инвертированных файлах поиск отбираемых по этим параметрам спектров реализуют методом прямого доступа и на современных компьютерах практически мгновенно.

Второй способ предварительного отбора - сопоставление спектров не в их полном объеме, а лишь по определенным спектральным признакам. В случае дискретного описания спектров в качестве параметра может выступать присутствие в спектре пробы одного или нескольких наиболее интенсивных пиков данного эталона, или наиболее информативных пиков, и т.п. [93, 99]. При сопоставлении по нескольким признакам можно потребовать наличия некоторой части пиков пробы в эталонном спектре, например трех из пяти наиболее интенсивных.

Наконец, для предварительного отбора можно использовать спектры одного вида (например, ИК) или хроматографические данные, а для окончательной идентификации - спектры другого вида (например, МС). Рассмотрим некоторые приемы предварительного отбора для разных спектральных методов.

Масс-спектрометрия. Требование наличия в спектре пробы 6 пиков из 8 наиболее интенсивных пиков проверяемого эталонного спектра приводит к сокращению достаточно крупных БД до списков, содержащих от 50 до 1000 эталонных спектров. Автоматизированный выбор ограниченного числа пиков, по которым ведется предварительный отбор, обладает рядом недостатков, например, отобранные пики могут оказаться сгруппированными в одной и малоинформативной области спектра. Особенно это характерно для случая масс-спектрометрии. Поэтому для целей предварительного отбора стремятся отобрать наиболее значимые пики, расположенные равномерно по всему спектру.

Высокую селективность отбора наряду с отмеченными выше обеспечивают параметры, описывающие совокупность спектральных данных, например “индекс смещения ионной серии”, “словарь масс-спектров” [100] и др. [91-92, 99-102]. При использовании метода хроматомасс-спектрометрии эффективным способом предварительного отбора является проверка по индексу удерживания, если соответствующая информация содержится в базе данных.

Предварительный отбор, проводимый по совокупности таких признаков, как молекулярная масса, *m/z* и интенсивности нескольких пиков; элементный состав и массы первичных нейтральных потерь, весьма эффективен и легко воспринимаем аналитиком. По-видимому, это и послужило причиной применения такой процедуры в ряде систем, например MSSS [103]. Особенность последней системы состоит в предоставлении аналитику права выбора признаков и условий отбора спектров. В этом ее достоинство и одновременно

недостаток; опытный спектроскопист быстро достигает требуемого результата, неопытный - увы... Поэтому более распространены системы, в которых критерии предварительного отбора выбраны разработчиками и не могут быть изменены пользователем.

ИК спектроскопия. Современные БД содержат наряду с полным описанием кривых поглощения так называемые сокращенные версии ИК спектров. В качестве последних обычно выступает дескрипторное описание (поисковый образ) – частоты и интенсивности наиболее значимых пиков, иногда и их полуширины. Это позволяет вести, как и в случае масс-спектрометрии, предварительный отбор по наиболее значимым или наиболее информативным пикам. Различие состоит в том, что в этом случае требуют не точного совпадения частот пиков пробы и эталона, а совпадение их с точностью до некоторого интервала частот. При решении идентификационных задач этот интервал обычно не превосходит 10-15 см$^{-1}$.

Спектроскопия ЯМР. Возможность дискретного описания спектров ^{13}C-ЯМР (диапазон химических сдвигов более 250 м.д. при типичной полуширине сигналов менее 0.01 м.д.) позволяет реализовывать достаточно быстрый и эффективный отбор. В качестве параметров предварительного отбора могут выступать: элементный состав соединения, число пиков спектра, наиболее характерные из них с указанием относительной интенсивности и положения в шкале м.д. Инвертирование БД по положениям пиков в шкале м.д., интенсивностям и/или мультиплетностям сигналов или наиболее информативным пикам обеспечивает прямой доступ и быстрый отбор требуемых записей.

Особенности спектроскопии ^{1}H-ЯМР обуславливают более сдержанное отношение к соответствующим БД и компьютерной идентификации соединений с их помощью. Как было показано в разделе 1.1.4, в ближайшие годы, вероятно, для решения поисковых задач будут широко использоваться не БД полных спектров ^{1}H-ЯМР, а версии БД вида “молекулярный граф – отнесения спектральных признаков к вершинам графа”**.** В этом случае предварительный отбор записей из БД реализуется на принципах, аналогичных спектроскопии ^{13}C-ЯМР.

1.2.2. Алгоритмы поиска информации в спектральных БД

При сопоставлении спектра пробы и эталонных спектров соединения возможны алгоритмы “прямого” и “обратного” поиска, а также комбинации этих алгоритмов. Различие двух подходов состоит в следующем.

При прямом поиске спектр пробы (опознаваемого соединения) поочередно сопоставляется с эталонными спектрами, а при подсчете степени совпадения по каждому эталону учитывают все сигналы (пики), присутствующие в спектре пробы. Степень

совпадения снижается при любом несовпадении признаков, например, при проявлении в спектре пробы лишних пиков по сравнению с эталонным спектром. При обратном поиске, наоборот, эталонные спектры сопоставляют со спектром пробы. В этом случае спектральные признаки, отсутствующие у проверяемого эталона, не учитываются, лишние пики в спектре пробы не снижают степени совпадения с данным эталоном (или, в терминологии [10] - не снижают сигнал присутствия проверяемого компонента).

В качестве примера можно привести многочисленные алгоритмы прямого [6], обратного [104] и комбинированного [90] поиска, разработанные группой Мак-Лафферти (F.W. McLafferty) для расшифровки масс-спектров. Достоинства и недостатки этих подходов обсуждены в литературе [105,106]. Показано, что при использовании прямого поиска опознание чистых веществ идет быстрее и нередко дает более однозначные результаты. Тем не менее на практике предпочтительны алгоритмы обратного поиска. Почему это так? Слишком часто в практике встречаются ситуации, когда анализируют недостаточно чистые вещества (например, не до конца разделенные при хроматографировании), и в спектре пробы (или хроматографической фракции), кроме пиков основного компонента, проявляются пики примесей. Это ухудшает результат прямого поиска основного компонента, но не снижает сигнал присутствия этого соединения при использовании алгоритмов обратного поиска (не случайно обратный поиск является основным приемом в компьютерном анализе смесей [10]).

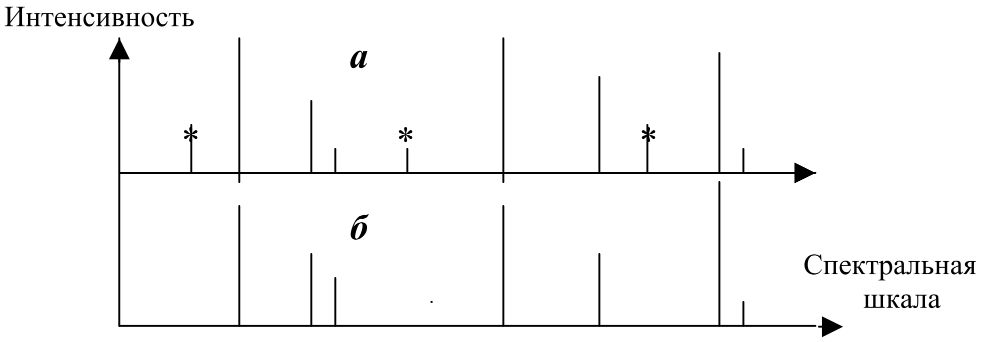


Рис. 1.3. Иллюстрация различий при сопоставлении двух спектров

(рисунок есть и в виде отдельного файла рис1-3)

Рисунок 1.3 схематически иллюстрирует сопоставление спектра некоторого соединения, содержащего примеси, с эталонным спектром БД. Сигналы примеси помечены звездочкой. Допустим, что представленный на рис.1.3. спектр пробы ***а*** сравнивают с эталонным спектром **б,** подсчитывая лишь *число пиков, совпадающих по своему положению в некоторой шкале* (бинарная или однобитовая кодировка). В случае прямого поиска окажется,

что спектр пробы не полностью совпал с эталонным (7 пиков совпали, а 3 не совпали). В случае обратного поиска будет отмечено, что все пики эталонного спектра проявились в спектре пробы, то есть спектры совпали полностью. Разница между двумя методологиями из этого примера очевидна. Если принять, что в БД содержатся именно спектры индивидуальных веществ, а примеси могут проявиться только в спектре пробы, то алгоритм “обратного” поиска получает явные преимущества.

Ряд ИПС позволяет реализовывать не только обратный или прямой виды поиска, но и их комбинацию – сочетание обоих видов. Этот подход оказывается еще более эффективным, так как он одновременно компенсирует (разумеется. лишь частично) как экспериментальные погрешности спектра пробы, так и возможные погрешности спектров БД.

Приведенный пример весьма идеализирован и далек от реальной практики, в которой при оценке сходства спектров важную роль играют не только положения пиков, но и относительные интенсивности, полуширина и другие параметры этих пиков. Однако в масс-спектрометрии сначала, как и в рассмотренном примере, проверяют совпадение пиков в шкале целочисленных значений массовых чисел. При сопоставления ИК спектров начинают поиск с проверки совпадения полос (с точностью до некоторой заранее заданной погрешности в шкале частот) и лишь затем учитывают остальные признаки спектра. Аналогично поступают в ЯМР-спектроскопии, проверяя совпадение сигналов пробы и эталона с точностью до заданной величины химического сдвига в шкале м.д.

На заключительном этапе сопоставления спектров пробы и эталонов рассчитывают количественные оценки степени совпадения (несовпадения) этих спектров в рамках выбранного алгоритма (для прямого или обратного поиска). К сожалению, терминология еще не устоялась: как синонимические, разными авторами используются понятия: *степень совпадения, критерий подобия, индекс подобия, степень сходства, мера близости, фактор совпадения и др.* Часто величину этого параметра нормируют; в этом случае его максимальное значение, характеризующее тождественность спектров пробы и эталона, принимают равным единице. Степень совпадения спектров характеризуют, например, коэффициентом линейной [107] или ранговой [108] корреляции, либо определяют долю площади перекрывания после наложения нормированных спектров, либо используют иные алгоритмы. Иногда рассчитывают *степень несовпадения спектров.* Ее величину для спектров, выраженных векторами в многомерном пространстве, часто характеризуют абсолютным или эвклидовым расстоянием. Выбор и способ расчета параметров совпадения зависят от вида сопоставляемых спектров и полноты их описания в соответствующей БД. В качестве примера в таблице 1.3 приведены некоторые из функций, используемых при сопоставлении масс-спектров.

При сравнении спектров, интенсивности пиков которых представлены бинарным кодом, используют логические операторы. Они позволяют быстро провести вычисление критериев, подобных *C*. В работах [111,112] показано, что результат поиска заметно изменяется в лучшую сторону при учете не только совпадающих, но и отсутствующих в сравниваемых спектрах пиков, а наиболее удовлетворительное решение достигается при значениях коэффициента μ в выражении для *C*, равного 2-3. Это справедливо как для случая масс-, так и для случая ИК спектроскопии молекул.

Таблица 1.3

Количественные оценки подобия (различия) сопоставляемых спектров

Параметр	Примечание
$D_1 = \Sigma \mid I_l - I_x \mid$, где $\Sigma I_l, \Sigma I_x = 1$	I_k , I_x – относительные интенсивности пиков с равными значениями *m/z*.
$D_2 = \Sigma \mid I_l^{1/2} - I_x^{1/2} \mid$, $\Sigma I_l^{1/2}, \Sigma I_x^{1/2} = 1$	
$D_3 = \Sigma (I_l - I_x)^2$, $\Sigma I_l^2, \Sigma I_x^2 = 1$	
$D_4 = \Sigma \mid I_l - I_x \mid$, где max $(I_l , I_x) = 100$	
$D_3 = \Sigma (cI_l - I_x)^2$, где $c = \Sigma I_l I_x / \Sigma (I_l)^2$	
$C_1 = \Sigma \mid L_l - L_x \mid$	L_l , L_x – или коды интенсивностей пиков (целые числа), или интенсивности, нормированные к максимальному.
$C_2 = \Sigma (L_l - L_x)^2 / \Sigma (L_l^2 + L_x^2)$	
$C_3 = \Sigma \mid L_l - L_x \mid / \Sigma (L_l + L_x)$	
$C = \mu N + \Sigma$ [(XOR) - μ(AND)], μ =2 - 3	*N* – число пиков в спектре исследуемого соединения, XOR и AND – сложение и умножение бинарных кодов, описывающих пик.
$P_1 = a/n$; $P_2 = 1/R\, \Sigma a_r/n$	*a* и n – число совпавших и сравниваемых пиков, *R*- число анализируемых массовых интервалов, *i* и *j* – номера совпавших по значениям *m/z* пиков в ряду пиков, ранжированных по интенсивности.

$P_3 = 1/n^2 \Sigma (n - \| i - j\|)$	
$P_4 = 1/R \Sigma 1/n^2 \Sigma (n - \| i - j\|)$	
SI = $\Sigma (i\ i)^b \Sigma\Sigma (i\ j)^b$	i и j – позиции элементов векторов, описывающих спектры неизвестного и БД; b – весовой коэффициент [109].
$MF10 = 1000 \Sigma(I_i/I_j)(I_i+I_j) / [\Sigma I_i + \Sigma I_j]$	I_i , I_j - относительные интенсивности пиков в сравниваемых спектрах [110].

Как видно из табл.1.3, в большинстве поисковых алгоритмов при сравнении спектров подсчитывают не только совпадения пиков по их положению, но и оценивают тем или иным способом совпадение (или несовпадение) их интенсивностей (например, функции D_1-D_4 в таблице). Попытки использования косвенной информации об интенсивностях, например, номера пиков в ранжированном ряду уменьшения их интенсивностей (P_3, P_4, S_I) оказались неудачными и не нашли применения в коммерчески распространяемых системах. Та же судьба постигла алгоритмы, требующие излишне строгого совпадения интенсивностей, поскольку этот поисковый признак в гораздо большей степени, чем положение пика, чувствителен к варьированию условий регистрации.

В масс-спектрометрических ИПС (как и в ИПС по другим видам спектроскопии) при оценке степени совпадения (различия) сравниваемых спектров вычисляют или абсолютные разности интенсивностей ($D_{1\text{-}4}$, C_2 в табл.), или величины, характеризующие отношение интенсивности сравниваемых пиков (см. C_3, C_4, MF_{10}). При этом зачастую используют приемы, близкие к традиционным методам “ручного” поиска; при оценке степени совпадения спектров предпочтение отдают совпадению наиболее интенсивных пиков. Близкие по характеру алгоритмы используют и в тех случаях, когда интенсивности представлены в огрубленной форме, например, в виде небольших целых чисел [94].

Наряду с рассмотренными эмпирическими и полуэмпирическими параметрами близости (различия) спектров в практике анализа широко используют ИПС, алгоритмы которых опираются на теорию информации и статистики. Наибольший интерес представляет расчет “confidence value” (K) [104] и его модификации [90], учитывающие статистическую значимость совпадения массовых чисел и интенсивностей в сопоставляемых спектрах. Критерий K вычисляется суммированием индивидуальных значений K_j, характеризующих значимость совпадения отдельных пиков в сравниваемых спектрах:

$$K_j = A_j + U_j + W_j - D,$$

где U_j - вклад “уникальности” значения m/z пика j, A_j - вклад значения интенсивности пика j

спектра БД, W_j - “фактор окна”, определяющий требуемую степень совпадения интенсивностей пиков анализируемого спектра и базы данных, D – “фактор разбавления”, определяющий общее уменьшение интенсивностей анализируемого спектра (заданной компоненты) вследствие возможного наличия в образце других примесей. Для спектра индивидуального соединения $D = 0$. Значения U_j рассчитывают из вероятности (P_j) появления пика с величиной m/z, равной m/z_j, в спектрах БД: $P_j=(1/2)^{Uj}$. Аналогично рассчитывают величину параметра A: $P_l = (1/2)^{Al}$, здесь P_l – вероятность появления в спектрах БД пиков с данным значением интенсивности. Для простоты расчетов при оценке значимости совпадений оперируют с целочисленными значениями величин. Этот подход используется в широко распространенной поисковой системе PBM, созданной группой Мак-Лафферти, и оказался весьма плодотворным при идентификации соединений по данным хроматомасс-спектрометрии. На принципах, учитывающих информативность положения и интенсивности сравниваемых пиков на основе уравнения Шеннона, разработан и алгоритм сопоставления спектров, используемый в системе “Компас-МС” [96].

В целом подход, базирующийся на учете информативности (характеристичности) идентификационных признаков, кажется вполне приемлемым и для других методов анализа [113]. Действительно, наиболее значимыми при сравнении спектров становятся в этом случае наиболее уникальные признаки, т.е. такие, которые редко встречаются в спектрах БД. Следовательно, если эти признаки устойчиво проявляются в спектрах тождественных соединений вне зависимости от условий регистрации, то этот факт нельзя объяснить случайными совпадениями. По наиболее информативным (характеристическим) признакам можно провести не только быстрый и эффективный предварительный отбор, но и окончательное ранжирование эталонных спектров. Важно, что подобный подход позволяет количественно оценить теоретическую достоверность получаемого результата поиска.

Рассмотрим в качестве примера упрощенный случай: $K=\Sigma K_j$, где $K_j=U_j+A_j$. Если в результате сравнения некоторого спектра со спектрами БД, насчитывающей 2^{15} спектров, найден спектр соединения, для которого $K=15$, то достоверность идентичности этого соединения неизвестному составляет 50%, поскольку среди 2^{15} спектров можно случайным образом найти еще один спектр с $K=15$. В случае $K=20$ – достоверность идентификации близка к 97%, так как вероятность случайного нахождения еще одного спектра с таким же значением К составляет 0.5^5. Напомним, однако, что в основу расчетов информативности положено предположение о независимости событий (появление данного пика с данным значением интенсивности вне связи с другими пиками спектра). Это предположение, как правило, не реализуется в случае спектроскопии молекул; при сопоставлении эталонных спектров можно убедиться, что пики в масс-спектрометрии, полосы поглощения в ИК,

сигналы ЯМР часто коррелируют друг с другом. Тем не менее в практике анализа информационные системы, построенные с учетом вероятности случайных совпадений и характеристичности признаков, весьма распространены. Этот подход наиболее последовательно реализован в ИПС, работающих со спектрами низкотемпературной люминесценции смесей полиаренов (см. главу 4).

Дескрипторное описание спектров позволяет представить их в виде векторов в d – мерном пространстве. Так, если X – совокупность массовых чисел m/z^x, далее для простоты m^x, и интенсивностей I_m^x спектра анализируемого соединения, а R – такая же совокупность m^r и I_m^r спектра из БД, то каждая из них может быть представлена векторами $X = (I_1^x, I_2^x,...,I_m^x)$ и $R = (I_1^r, I_2^r,...,I_m^r)$. При этом степень совпадения спектров (индекс взаимной корреляции) пропорциональна квадрату косинуса угла между векторами

$$C=k \cos^2 \alpha = k\ (XR)/(X)^2\ (R)^2,$$

где k – нормировочный множитель, $0 \le C \le k$, (XR) – скалярное произведение векторов.

По определению $(X)^2 = \Sigma\ (I_m^x)^2$; $(R)^2 = \Sigma\ (I_m^r)^2$; $(XR) = \Sigma\ (I_m^x I_m^r)^2$.

Если при вычислении $(R)^2$ учитывать только те составляющие вектора R, которые присутствуют в спектре X (режим обратного поиска), то получим величину $(R^*)^2 \le (R)^2$. Учитывая при расчете $(X)^2$ только пики, присутствующие в R – спектре, получим $(X^*)^2 \le (X)^2$. Подставляя $(X^*)^2$ и $(R^*)^2$ в выражение для индекса корреляции спектров, получим параметры $C^r \ge C$ и $C^x \ge C$, по максимальным значениям которых можно ранжировать поисковый ответ.

Общая черта многих поисковых алгоритмов – лежащий в их основе эмпирический характер рассчитываемых параметров совпадения (или различия) сопоставляемых спектров. Причина этого кроется в попытках преодолеть природу методов, обуславливающую нестабильность спектральных параметров и их связь с условиями регистрации спектров. В ряде случаев это приводит к достаточно сложным схемам расчета. Так, например, сравнение масс-спектров в работе [110] проводится с учетом 10 различных факторов:

$$MF=\Sigma\ k_i\ (MF_i)\ /\ 12,$$

где k_i – эмпирический коэффициент индивидуального фактора MF_i. В работах [87, 114-115] отбор и ранжирование спектров осуществляется с учетом шести факторов и эмпирически подобранных к ним весовых функций. В работе [116] на примере ИК спектроскопии убедительно показано, что результат сопоставления полных ИК спектров по Евклидовой метрике, уступает таковому, получаемому с помощью эмпирического алгоритма, как при сопоставлении полных спектров, так и при их представлении в дескрипторном описании.

Для оценки меры близости ИК спектров также используют разные математические функции. Среди них: сумма квадратов разностей; сумма абсолютных величин разностей интенсивностей спектральных полос или их первых производных [117]; вычисление

коэффициентов корреляции [118] или функции взаимной корреляции [119-120]; нечетные моменты в функции взаимной корреляции [121], метрика Гротча [111], методы нечёткой логики [120, 122-123] и др. [43, 124]. Типичные алгоритмы измерения спектрального расстояния приведены в работе [125]. Хорошие результаты достигаются и при использовании эмпирических функций [126], что, несомненно, обусловлено высокой зависимостью спектра к условиям регистрации. Подчеркнем, что ИК спектры тождественных соединений, строго говоря, не тождественны, хотя бытует мнение, что "ИК спектр – своеобразный паспорт соединения" или его "отпечаток пальцев".

Для ускорения процедуры поиска в ИК спектроскопии используют различные приемы, например, обнуление малоинформативных спектральных интервалов [127] или файлы сокращенных спектров [128]. Метод поиска, в котором для идентификации спектров используют только фазовые компоненты Фурье-преобразования [129], позволяет идентифицировать спектры, полученные в разных инструментальных условиях. Ряд эффективных приемов на основе методов факторного анализа и собственных векторов предложен в работах [130-132]. Построение БД в виде иерархических деревьев использовано в [133]. В этом случае затраты на построение дерева окупаются быстротой поиска. Разработан алгоритм, позволяющий повторять поиск в узлах дерева, что приводит к росту достоверности результата идентификации, но требует больше времени [134]. Предварительный отбор спектров по положению полос с последующим сравнением полных спектров выборки [135] и оценка подобия отдельных частей полных спектров рассмотрены в работе [136].

1.2.3. Результат поиска

При составлении ответа на запрос пользователя все ранее отобранные в БД эталонные спектры ранжируются по величине параметра, который в рамках принятого алгоритма характеризует степень их подобия спектру пробы. Пользователь получает перечень эталонов, занимающих первые места в ранжированном списке (hit list). При этом возможны два способа ограничения перечня: либо в машинный ответ включается постоянное и заранее заданное число эталонов (например, первые 10), либо в ответ включаются все эталоны, у которых степень совпадения со спектром пробы оказалась выше определенного критического значения. Критерии отбора (критические значения ранжирующего параметра) обычно устанавливаются эмпирически. Вопрос о выборе критериев отбора особенно важен в случае анализа смесей [113,137], поэтому он специально будет обсуждаться в разделе 4.3. Однако ни тот, ни другой способ не гарантирует получения правильного и однозначного ответа.

Влияние условий регистрации, способа описания спектра и алгоритма сопоставления часто приводит к тому, что искомый спектр (и соединение) могут оказаться не на первом месте в ранжированном списке спектров, отобранных из БД. Поэтому результат поиска неоднозначен, машинный ответ представляет собой лишь *перечень возможных вариантов идентификации*. Окончательное заключение обычно выносит сам исследователь, визуально сравнивая отобранные эталонные спектры со спектром пробы. Это связано с несовершенством компьютерных алгоритмов сопоставления спектров, пока что они менее чувствительны к тонким различиям спектров, которые может принять во внимание исследователь (форма пиков и т.п.). Желательно использовать все имеющиеся сведения о пробе (не указанные в запросе), а также логически проанализировать машинный ответ. Например, скачкообразное изменение в величинах параметра ранжирования у разных эталонных спектров в поисковом ответе служит основанием для более внимательного анализа записей, расположенных выше наблюдаемого “скачка” или излома.

Правильность машинного ответа оценивают по результатам контрольных поисков, выполняемых по реальным спектрам таких соединений, спектры которых (обычно записанные в несколько иных условиях регистрации) имеются в базе данных. Для этой же цели используют и более упрощенный прием: поиск тождественных соединений, если в БД одно и то же соединение представлено двумя и более записями. Идентифицирующую способность характеризуют чаще всего величиной, равной отношению числа положительных решений (N^{+}) задач идентификации к общему числу контрольных поисков (N):

$P=N^{+}/N$, %.

Положительным решением считают то, в котором “неизвестное” соединение оказывается на первом месте машинного ответа, или входит в первую тройку, или в первую пятерку и т.п. Наибольший интерес представляет оценка, отнесенная к первому из отбираемых из БД соединений. Соответствующие результаты представлены в табл.1.4. Данные получены при проверке ряда масс-спектрометрических ИПС, созданных в 80-ые годы.

Если поиск ведется только с использованием масс-спектральных данных (*m/z, J*), то вероятность правильной идентификации единичного соединения по первому отобранному из БД спектру колеблется в области от 70 до 80%. Сходные результаты получают и в более поздних версиях поисковых систем по масс-спектроскопии, все шире используемых в исследовательской и аналитической практике. Аналогичны по своей эффективности и ИПС по ИК спектроскопии. В случае ^{13}C-ЯМР вероятность правильной идентификации соединения по первому отобранному из БД спектру обычно более высокая.

Таблица 1.4

Результаты проверки идентифицирующей способности некоторых масс-спектрометрических ИПС

Число спектров в БД	Число спектров “неизвестных” соединений	Доля правильных решений, %
8000	239	80.0 [а]
6880	125	90.4 [б]
6000	85	70.0 (87) [а]
6652	80	76.0 [б]
17000	43	78.0 [в]
7600	32	90.6
13000	28	92.8

а – сопоставлялись спектры только таких соединений, которые отличаются от “неизвестного” по молекулярной массе на 3 а.е.м. б – решение считалось правильным и в тех случаях, когда на первом месте ответа ИПС находился изомер “неизвестного”. в – исследовались спектры только углеводородов. Правильным решением во всех случаях считалось появление опознаваемого соединения на первом месте машинного ответа.

Вероятность появления искомого соединения на первом месте машинного ответа увеличивается, если воспользоваться априорной информацией о соединении, например, указав в запросе наряду со спектральными данными молекулярную массу (ММ) соединения. Это традиционный прием большинства пользователей. Но в нем, как показано в работах [138-139], содержится методологическая неточность. При задании в запросе на поиск молекулярной массы сокращается лишь число просматриваемых в БД записей, но это не гарантирует однозначную идентификацию. Вероятность правильной масс-спектрометрической идентификации по первому соединению из поискового ответа в этом случае составляет от 80 до 90%. Если же дополнительной информацией воспользоваться не на стадии формирования запроса, а при анализе результата поиска, проведенного только по спектральным данным, то в ряде случаев число возможных гипотез о строении неизвестного

существенно сокращается. Как показывает практика, число гипотез о строении неизвестного, если исходный поисковый ответ содержит 10 записей, часто уменьшается до одной, соответствующей искомому соединению. Если все же ответ и в этом случае содержит несколько записей, то данный прием, как правило, приводит к увеличению “скачка” в величинах параметров совпадения (различия) отбираемых спектров.

В таблице 1.5. в качестве примера приведены два поисковых ответа, полученные по масс-спектру соединения с предполагаемой молекулярной массой, равной 244 а.е.м. В первом случае в поисковом запросе молекулярная масса как поисковый признак не задана. Поиск проводится только по спектральным параметрам (*m/z* и интенсивности пиков). Во втором – поиск в БД проведен среди соединений, обладающих данным значением ММ.

Таблица 1.5.

Примеры поисковых ответов

ФС	Название соединения	ММ	МФ
Ответ 1			
66	Lauric acid-mono-TMS	272	$C_{15}H_{32}O_2Si_1$
64	Capric acid-mono-TMS	244	$C_{13}H_{28}O_2Si_1$
64	Myristic acid-mono-TMS	300	$C_{17}H_{36}O_2Si_1$
64	Palmitic acid-mono-TMS	328	$C_{19}H_{40}O_2Si_1$
63	Monotrimethylsilyl myristic acid	300	$C_{17}H_{36}O_2Si_1$
63	Heptadecanoic acid-mono-TMS	342	$C_{20}H_{44}O_2Si_1$
62	Octanoic acid-mono-TMS	216	$C_{11}H_{24}O_2Si_1$
59	Monotrimethylsilyl lauric acid	272	$C_{15}H_{32}O_2Si_1$
59	Monotrimethylsilyl palmitic acid	318	$C_{19}H_{40}O_2Si_1$
55	Stearic acid-mono-TMS	356	$C_{21}H_{44}O_2Si_1$
Ответ 2			
64	Capric acid-mono-TMS	244	$C_{13}H_{28}O_2Si_1$
50	Decanoic acid-mono-TMS	244	$C_{13}H_{28}O_2Si_1$
37	Desoxypodocarpinol	244	$C_{17}H_{24}O_1$
37	1-Ethyl-*trans*-3-N-heptyl-(2,3-dihydroinden)	244	$C_{18}H_{28}$
33	Podocapra-8,11,13-trien-12-ol	244	$C_{17}H_{24}O_1$
32	2,6-*di*-*tert*-Butyl-1,2,3,4-tetrahydronaphthalene	244	$C_{18}H_{28}$
30	1-Ethyl-*cis*-3-N-heptyl-(2,3-dihydroinden)	244	$C_{18}H_{28}$
30	2,7-*di*-*tert*-Butyl-1,2,3,4-tetrahydronaphthalene	244	$C_{18}H_{28}$

30	5-Propionyl-6-methyl-4-phenyl-1,2,3,4-tetrahydro-2-oxopyrimidine	244	$C_{14}H_{16}O_2N_2$
28	Ethyl-2,4-diethyl-4-phenylbutadien-2,3-oate	244	$C_{16}H_{20}O_2$

Легко видеть, что как в первом, так и во втором случаях невозможно принять однозначное заключение о строении исследуемого соединения. Факторы совпадения спектров из БД с предъявленным спектром (в таблице обозначены как ФС.) достаточно равномерно убывают в обоих случаях. Обратим, однако, внимание на то, что если информацией о величине ММ соединения воспользоваться не при задании поискового запроса, как это часто делают (см. ответ 2), а при анализе результата поиска, проведенного только по спектральным данным (см. ответ 1), то число гипотез о строении неизвестного сразу уменьшается с десяти до одной (только одно соединение имеет ММ равное 244).

При оценке идентифицирующей способности можно использовать величину P_n, характеризующую способность ИПС поставить идентифицируемое соединение среди n первых эталонов, отобранных из БД (например, n = 3, 5, 10 и т.п., [30,140,141]). Так, в первую тройку опознаваемое соединение обычно входит с частотой 90 % и выше [142].

Возможны и другие приемы оценки результата идентификации. Например, в работе [143, 144 и др.] для этой цели используют параметр

$RL = 100\, I_c / (I_c + I_f)$.

Здесь *RL (reliability)* – достоверность принятия решения об идентификации объектов контрольной выборки, I_c и I_f – число корректных и ошибочных идентификаций при заданном пороговом значении параметра близости (или различия) спектра запроса от спектров БД. Иная количественная оценка эффективности библиотечного поиска предложена в [145-146]. Используя её, в работе [145] изучали влияние привнесенного в ИК спектры шума на эффективность поиска и установили, что для спектров в газовой фазе отношение сигнал/шум от 2 до 5 приводит к хорошим, а свыше 5 - к отличным результатам поиска. Мера количественной надежности (достоверности) поисковых результатов использована в [147] для сопоставления двух метрик сравнения спектров (евклидовой и скалярного произведения векторов).

Обычно пользователь располагает сведениями об элементном составе или молекулярной массе опознаваемого соединения, полученными традиционными методами: CHN- анализ, эбулиоскопия, криоскопия, осмометрия и т.п. Однако в некоторых случаях эти данные неизвестны (например, при опознании хроматографических фракций). Программное обеспечение современных ИПС позволяет получить эти данные из масс-спектра пробы, одновременно с решением идентификационной задачи. Разумеется, при определении молекулярной массы или брутто-формулы соединения по его спектру, как и в случае

анализа поискового ответа, надо помнить о вероятностном характере достигаемого результата. Как показано в [148]), правильные значения ММ индивидуального соединения содержатся среди первых 10 значений, предлагаемых ЭВМ, с вероятностью, близкой к 98% (даже в случаях отсутствия в спектре пика молекулярных ионов [149] или при наличии примесей [150]). Более скромные результаты получены при определении элементного состава — 86% среди первых 10, генерируемых компьютером [148]. Получаемую информацию можно использовать и при анализе поискового ответа при решении задачи идентификации соединения.

Более сложны приемы оценки эффективности поиска не полностью тождественных эталонам соединений, а подобных им по строению. В этом случае возникает проблема определения степени структурного подобия соединений. Она неоднократно обсуждалась в литературе [151]. Удачный пример ее решения представлен работой [116], которую стоит сравнить с данными [152].

Использование дополнительной информации на этапе анализа машинного ответа, полученного только по набору данных, характеризующих спектр (m/z-интенсивности пиков или частота-интенсивность поглощения), во многих случаях позволяет решить и вопрос о том, содержится ли вообще спектр изучаемого соединения в базе данных. Проверить наличие или отсутствие спектра соединения в БД можно и другими средствами, например, по брутто-формуле соединения, его названию или регистрационному номеру. Современные ИПС, как правило, содержат информацию о кодах структурной формулы соединений и средства их поиска в БД, что также можно использовать при проверке содержимого БД.

1.2.4. Оценка эффективности ИПС и приемы ее повышения

Из всего вышеизложенного видно, что структура и функциональная схема автоматизированных систем информационного поиска, реализуемые с помощью современных персональных компьютеров, во многом повторяют черты соответствующих неавтоматизированных процедур. Основное различие заключается в том, что пользователь автоматизированной системы отделен (или удален) от процедуры поиска. Механизм поиска часто выступает для него своеобразным “черным ящиком”, а результат полностью определяется опытом и “искусством” составления запроса. Поэтому один из важнейших принципов создания информационно-поисковых систем (ИПС) состоит в разработке таких баз данных (БД), такого аппарата поиска и языка общения с системой, которые обеспечивали бы простоту достижение того же самого результата, который мог бы быть получен пользователем без обращения к системе, путем тщательного анализа всех архивных данных, относящихся к интересующему вопросу.

Эффективность использования ИПС напрямую зависит от темпов наполнения соответствующих БД новой информацией и решения связанных с этим организационных, технических и коммерческих вопросов. Некоторое представление об этом дает приведенная на рис. 1.4 упрощенная схема взаимодействия (с использованием сети INTERNET) поставщиков, продавцов и потребителей информации, хранящейся в удаленных БД коллективного пользования.

На основе первичной информации (книги, журналы, тематические БД и т.п.) производитель создает базу данных и передает ее оптовому продавцу (вендору). “Продавец” выполняет следующие основные функции. Преобразует все БД, созданные разными производителями, так, что их дальнейшее использование становится возможным на едином языке. Хранит и пополняет БД, организует доступ к ним через коммуникационные линии связи, регулирует финансовые отношения с пользователями и поставщиками информации. “Посредник” в данной схеме – это специалист, обладающий знаниями квалифицированного поиска релевантной запросам разнообразных пользователей информации. Очевидно, что пользователь может и сам обращаться к продавцам или производителям БД, что и реализуется на практике, особенно в случаях доступа к информации, выставленной в сети на некоммерческой (бесплатной) основе.

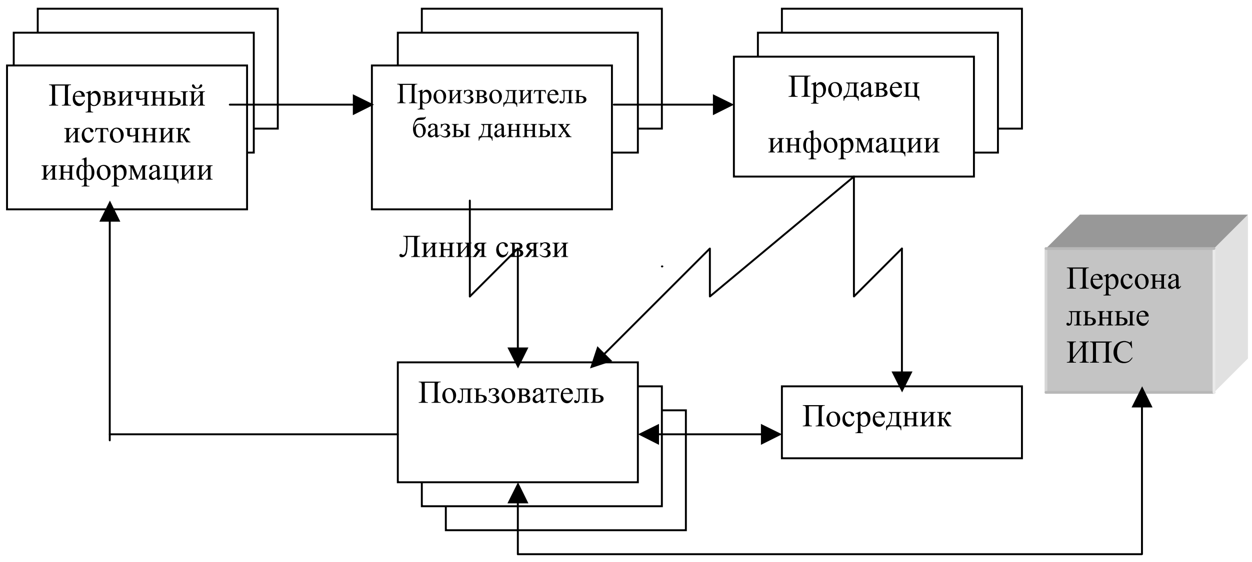


Рис.1.4. **Схема взаимодействия пользователя с БД в сети Интернет**
(есть в виде отдельного файла рис1-4)

Отметим. что до недавнего времени сложная схема взаимодействия поставщиков и потребителей спектральной и структурной информации с архивами приводила к тому, что значительная ее часть в силу трудности поиска в информационных потоках была полностью или частично утерянной и не использовалась в практике. Традиционный для химии прием

преодоления информационного кризиса путем создания обобщенных корреляционных таблиц или других обобщений вида “структура-свойство” оказался успешным лишь в некоторых случаях, например, в спектроскопии ядерного магнитного резонанса (ЯМР). В то же время другие данные, например, инфракрасной (ИК) и особенно масс-спектрометрии неудовлетворительно описываются подобными приемами, здесь БД и соответствующие им ИПС оказались незаменимыми.

Потребитель информации заинтересован в очень быстрой селекции из БД необходимых данных, что непросто сделать при детальном представлении в одной БД информации широкого профиля. Именно поэтому и создаются ИПС двух различных типов: а) ИПС коллективного пользования, доступные, например, по Международной научно-технической сети (STN International), и б) ИПС индивидуального (персонального) пользования, например, по масс-спектрометрии, поставляемые вместе с современным спектральным оборудованием.

ИПС коллективного пользования создают и устанавливают в крупных научных или коммерческих организациях, способных поддерживать и непрерывно пополнять БД, обеспечивать доступ многочисленных удаленных пользователей, решающих самые разнообразные поисковые задачи. Поставщики (продавцы) таких ИПС предоставляют пользователям не только БД (файлы) по различным областям знаний, но и наполняют их в ряде случаев качественно новой информацией. Разумеется, ИПС коллективного пользования обладают развитым диалогом, их стремятся сделать доступными пользователям любой области знания и различной квалификации. Последнее обстоятельство играет чрезвычайно важную роль, и давно осознано разработчиками систем. Система не должна “отпугивать” сложностью языка общения ее новых потребителей. Простота диалога с ИПС, вид выдачи искомых документов с целью легкого восприятия получаемого ответа особенно характерны для ИПС персонального назначения. Это способствует преодолению своеобразного “психологического барьера”, возникающего при обращении ко все еще нетрадиционным для аналитической химии методам поиска и анализа информации, рассчитанным на широкий круг потребителей.

ИПС индивидуального (персонального) назначения, в отличие от коллективного, используют специализированные проблемно-ориентированные базы, например: по биологически активным веществам определенных химических классов, по полимерам и продуктам их деструкции, по отходам конкретных видов производств и т.п. В случаях молекулярной спектроскопии – это обычно БД по масс-спектрометрии, ИК или УФ (ультрафиолетовой) спектроскопии, или спектроскопии ЯМР на определенных ядрах. Базы данных в этих случаях содержат полные спектральные образы или своеобразную выжимку

из информации основного хранилища. Характерная черта персональных систем – алгоритм поиска информации, как правило, жестко привязан к виду соответствующих фактографических данных, а сама поисковая процедура позволяет работать в режиме “on-line” с соответствующим спектральным инструментом. Потребители персональных ИПС – научно-исследовательские организации, аналитические лаборатории, лаборатории судмедэкспертизы, таможни, службы охраны окружающей среды, агрохимические службы и др. Число потребителей быстро увеличивается.

Базы данных персональной ИПС и ИПС коллективного пользования, как ясно из изложенного выше, могут значительно отличаться друг от друга. Так, например, спектральная часть баз данных по молекулярной спектроскопии в первом случае может содержать лишь наиболее характерные спектральные признаки (ограниченное число наиболее интенсивных пиков или полос поглощения), тогда как во втором – полное описание спектральных данных или вид соответствующей графической зависимости. Заметим, однако, что быстрый прогресс в области вычислительной техники приводит к частичному стиранию граней между базами данных для ИПС обоих типов. Тем не менее, по чисто экономическим соображениям вряд ли этот процесс завершится полным интегрированием.

Исследованию эффективности поиска в конкретных ИПС по ИК спектроскопии посвящены обстоятельные работы группы Клерка [153-154]. В частности, в [153] оценивается влияние условий регистрации образца (концентрации, примесей, коррекции фоновой линии) на результаты спектрального поиска в БД коммерческих ИПС, а в работе [154] сопоставляются результаты использования около 40 различных ИПС на одной и той же выборке “эталонных” данных. Тем не менее, по имеющимся литературе сведениям до сих пор не представляется возможным достаточно объективно судить о преимуществах или недостатках той или иной поисковой системы. В большинстве случаев их результативность оценивают при помощи различных БД и на различных тестовых выборках спектров “неизвестных” соединений. Сравнительная оценка различных поисковых алгоритмов (систем) в одинаковых экспериментальных условиях [48, 122-124] выполнена или для определенных классов соединений, или на сравнительно малых по объему БД [87, 116].

Выше были рассмотрены некоторые аспекты повышения эффективности использования ИПС при решении задачи идентификации соединения. Обмечено, в частности, что вероятность корректной идентификации определяется не только алгоритмом отбора данных из БД, способом их представления, объемом БД и качеством хранящихся записей. Важную роль, как показано, играют и другие известные о соединении сведения, например, данные о молекулярном весе (молекулярной массе) или элементном составе (молекулярной формуле). В этой связи дополнительно подчеркнем, что вероятность идентификации соединения средствами

ИПС резко возрастает, если анализируются результаты поисков, проведенных не по спектру одного из видов молекулярной спектроскопии, а по нескольким спектрам различной природы.

Наиболее простой путь идентификации соединения в этом случае заключается в пересечении результатов поисков, полученных с помощью различных ИПС (например, по ИК и масс-спектрометрии). При этом факт появления одного и того же соединения в ответах обеих ИПС может играть решающую роль при принятии решении об идентификации вне зависимости от места, занимаемого этим соединением в поисковых ответах. Приведем лишь один пример. При опознании 3-метокси-1-бутанола [155] искомое соединение находилось на седьмом месте поискового ответа по ИК спектроскопии и на десятом месте ответа ИПС по масс-спектрометрии. Перекрестный анализ обоих ответов приводит к единственному результату – соединению, идентичному “неизвестному”.

Рассмотренный прием, по-видимому, практически исключает и получение ошибочного результата идентификации. Так как очень мала вероятность случайного появления одного и того же соединения в ответах поисковых систем, оперирующих со спектрами различной физической природы. Косвенным подтверждением этого служит тот факт [155], что при решении многочисленных поисковых задач, как указывают авторы, ни разу не встретилась ситуация, когда одно и тоже соединение, будучи не идентичным определяемому, появилось бы одновременно в ответах двух различных ИПС.

Пока, к сожалению, прием использования нескольких ИПС (по различным видам спектроскопии) для решения задач идентификации соединения не находит должного применения. Две основные причины вызывают это. Первая – отсутствие “под руками” массового пользователя соответствующих программно-технических средств. Вторая – заключается в том, что даже широко используемые в практике соединения не всегда представлены всеми своими спектрами в соответствующих БД. Последнее обстоятельство определяет необходимость дальнейшего наращивания объема БД по всем видам молекулярной спектроскопии и, разумеется, другим разделам качественного органического анализа.

1.2.5. Использование масс-спектрометрических ИПС в анализе сложных объектов

Рассмотрим в качестве примера, как с помощью ИПС “Поиск-МС” был определен качественный состав достаточно сложного объекта - фенольной фракции препарата “Вахтоль” [156]. Исследование поставлено в связи с распространенностью способов бездымного копчения продуктов с помощью коптильных препаратов, изготавливаемых из древесных пиролизатов. Использование коптильных препаратов позволяет значительно снизить или полностью устранить содержание канцерогенных веществ в продуктах. Принято считать, что появление аромата копчености связано с попаданием в продукт соединений

фенольного характера.

Исследование вели методом хромато-масс-спектрометрии. Навеску препарата (около 100 г) подвергли шестикратной экстракции диэтиловым эфиром. Выход эфирорастворимой части после отгонки эфира составил 2,48% от исходного препарата. Эфирорастворимую часть авторы указанной работы разделили на кислотную, фенольную и нейтральную фракции. Анализ фенольной фракции проводили на хромато-масс-спектрометре “Hewlett-Pacard 5960B”, используя стеклянную капиллярную колонку типа WCOT(25x0,25 мм) с неподвижной жидкой фазой OV-101. Программированное изменение температуры от 70 до 270 ^{o}C осуществляли со скоростью 5^{o}C/мин. Хроматограммы регистрировали по полному ионному току. Масс-спектры записывали при энергии ионизирующих электронов 70 эВ. На рис.1.5. представлена хроматограмма фенольной фракции. Пики, для которых получены качественные масс-спектры, пронумерованы от 1 до 19.

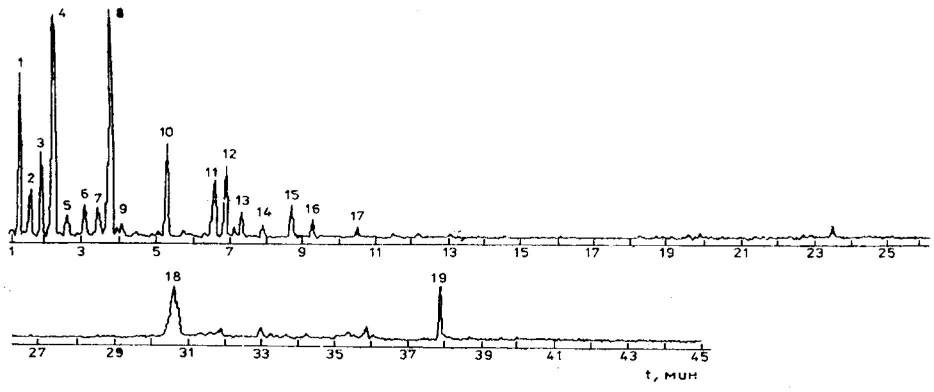


Рис.1.5. **Хроматограмма фенольной фракции коптильного препарата “Вахтоль”**

(есть в виде отдельного файла рис1-5)

Спектры поочередно анализировали на ЭВМ с помощью ИПС “Поиск-МС”, разработанную в НТЦ ХИ [155]. Поисковый ответ представляет собой список названий соединений БД, спектры которых наиболее подобны спектру анализируемого соединения, т.е. дают наивысшие значения фактора совпадения (ФС).

Таблица 1.5.

Машинный ответ (первые пять соединений) при опознании пика № 14

№	Мол. масса	Брутто-формула	Соединение	ФС
1	164	$C_{10}H_{12}O_2$	2-Метокси-4-(проп-2-енил)-фенол (изоэвгенол)	92
2	164	$C_{10}H_{12}O_2$	*Транс*-Изоэвгенол	89
3	164	$C_{10}H_{12}O_2$	2-Метокси-4-(проп-1енил)-фенол (эвгенол)	88
4	164	$C_{10}H_{12}O_2$	Эвгенол	86

5	164	$C_{10}H_{12}O_2$	*Транс*-Изоэвгенол	81

В качестве примера в табл.1.5. приведен ответ системы “Поиск-МС”, полученный при анализе масс-спектра одной из компонент (пик № 14) фенольной фракции. Указаны соединения с наиболее высокими значениями фактора совпадения. Как видно, на первом месте ответа ИПС находится изоэвгенол, масс-спектр которого наилучшим образом совпадает с анализируемым спектром (ФС=92%). Визуальное сравнение экспериментального и эталонного спектров также показало их хорошее совпадение, что и позволило идентифицировать соединение №14 как изоэвгенол. Аналогичным образом были опознаны еще 17 компонентов фенольной фракции препарата (табл.1.6.), причем отнесение ряда пиков подтверждается совпадением характеристик хроматографического удерживания. Результаты количественного анализа показывают, что основными компонентами фракции являются 4-метилгваякол, гваякол и фенол. На их долю приходится более половины содержания фенолов фракции.

Таблица 1.6.

Результаты качественного и количественного анализа фенольной фракции

№ пика	Соединение	Содержание,%
1	Фенол	8,96
2	Метилфурфуриловый спирт	3,99
3	о-Крезол	4,30
4	Гваякол	19,51
5	Этилфурфуриловый спирт	1,60
6	2,5-Диметилфенол	2,55
7	4-Этилфенол	2,59
8	4-Метилгваякол	23,22
9	Венилфениловый эфир	0,78
10	4-Этилгваякол	4,26
11	Сирингол	3,66
12	Эвгенол	3,33

13	Ванилин	1,43
14	Изоэвгенол	0,80
15	4-метилсирингол	1,91
16	Ацетованилон	1,05
17	2,6-Дигидрокси-4-метоксиацетофенон	0,82
18	Не идентифицировано	12,92
19	3,6-Дигидрокси-2,5-бис-(4-метоксифенил) - 1,4-бензохинон	2,30

Другим примером может быть иследование трансформации углеводородов (исходная смесь CH_4 – C_6H_6 – N_2O) на цеолите. Исследовали реакционную смесь, полученную на цеолите ZSM-5 при температуре 418^{o}C за 2000 с. (подробности см. в работе [157]. Используемое оборудование: хроматограф HP 6890A, снабженный масс-спектрометрическим детектором MSD 5972A, колонка 30м. HP-5MS. Зарегистрированная авторами хроматограмма реакционной смеси представляет собой набор хорошо разрешенных пиков (свыше 20). Были получены масс-спектры основных пиков. Учитывались также времена удерживания эталонных соединений. Использована коммерчески поставляемая база данных NIST (Национальное бюро стандартов и технологии, США). С помощью системы “Поиск-МС” при расшифровке масс-спектров получены следующие результаты отнесения хроматографических пиков: 1- гексан (растворитель), 2- бензол, 3- толуол, 4- *n*-октан, 5- этилбензол, 6- *m*+*p*-ксилол, 7- *o*-ксилол, 8- фенол, 9- индан, 10,11- крезолы, 12- нафталин, 13- 2-метилнафталин, 14- дифенилметан. Результаты опознания пиков по временам удерживания и по масс-спектрам (первое соединение в машинном ответе) в основном совпадают.

Заметим, что современная химическая литература пестрит сообщениями об успешном использовании методов хромато-масс-спектрометрии, применении баз данных и систем компьютерной идентификации для решения самых разнообразных научных и прикладных задач. Многолетний опыт работы НТЦ ХИ и других лабораторий, в которых эксплуатируются системы компьютерной идентификации веществ подтверждает важную роль информационных технологий (в сочетании с современной приборной базой спектрального анализа) в повышении эффективности качественного и количественного анализа.

Глава 2

КОМПЬЮТЕРНЫЕ СРЕДСТВА И МЕТОДЫ УСТАНОВЛЕНИЯ

СТРОЕНИЯ

НЕИЗВЕСТНОГО СОЕДИНЕНИЯ ПО СПЕКТРАЛЬНЫМ ДАННЫМ

В предыдущей главе показано, что в тех случаях, когда мы имеем дело с соединениями, спектры которых представлены в базах данных, они могут быть идентифицированы с помощью информационно-поисковых систем. При опознании соединений, спектры которых отсутствуют в БД, возникает более сложная и многоплановая задача - определение структурной формулы неизвестного соединения по спектральным данным. В общем случае эта задача не имеет однозначного решения и требует существенных затрат времени исследователя на анализ спектров, просмотр литературы, формулирование гипотез и т.п. При этом на конечный результат во многом влияет субъективный фактор, связанный с квалификацией исследователя и его опытом работы с различными видами спектров. Неудивительно поэтому, что для решения рассматриваемой задачи все чаще прибегают к помощи ЭВМ. Развитие работ в этом направлении привело к созданию самостоятельной области научных исследований, называемой в иностранной литературе "Computer-Assisted Structure Elucidation" [7].

Из двух основных путей установления структуры органических соединений по спектральным данным (метод искусственного интеллекта [158-161] и метод, основанный на применении фактографических БД [8,14,72, 162-163]) в данной главе будет рассмотрен лишь второй, поскольку системы на основе БД позволяют более объективно использовать накопленный экспериментальный материал и не имеют принципиальных ограничений на классы анализируемых объектов. Результативность решения задач в этом случае определяется наличием в используемых БД сравнительно небольшого числа соединений, подобных исследуемому как в спектральном, так и в структурном отношениях. Расширение баз данных (а это неизбежный и интенсивный процесс) способствует росту эффективности метода, не требуя дополнительных затрат на разработку новых алгоритмов и программ.

Ниже рассмотрены результаты исследований Научно-технического центра химической информатики (НТЦ ХИ) в области решения структурно-аналитических задач с применением различных видов молекулярной спектроскопии. Результаты получены с использованием БД, включающих ~70000 ИК спектров, ~50000 масс-спектров, ~44000 ^{1}H- и ~27000 ^{13}C-ЯМР спектров [47, 68]. Кроме спектральной информации, эти БД содержат сведения о структурных формулах соединений в виде двумерных химических графов. Вершины графа - структурные дескрипторы, а ребра - химические связи между ними (одинарная, двойная, тройная, "ароматическая", водородная и т.д.). В качестве структурных дескрипторов выступают отдельные атомы и микрофрагменты (CH_3, CH_2, CH, OH, CO, NO_2, SO_2 и т.п. -

всего 28 дескрипторов), наиболее часто используемые при описании структурных формул органических соединений. Кодовые записи структур содержат сведения о матрице связности химического графа, типах его вершин и координатах, позволяющих выводить из памяти ЭВМ структуру в привычном для химика виде.

Анализировали спектры “неизвестных” соединений, доступные большинству аналитиков, а именно: масс-спектры низкого разрешения; ИК спектры в диапазоне 4000-40 $см^{-1}$; спектры ^{1}H ЯМР (ПМР) с интегральными интенсивностями сигналов и спектры ^{13}C ЯМР, полученные в режимах полной и частичной развязки спин-спинового взаимодействия с ядрами ^{1}H.

2.1. Методология решения структурных задач с помощью БД

Из анализа литературы и накопленного опыта [68,164] просматривается возможность общего подхода к установлению строения химических соединений с помощью БД "структура-спектр". Этот подход предусматривает выполнение следующих основных операций:

– поиск в БД соединений, спектры которых наиболее близки предъявленному;
– анализ результатов поиска с целью получения информации об особенностях строения неизвестного соединения;
– анализ выявленной информации и генерирование на ее основе гипотез о строении этого соединения (списка структур);
– ранжирование полученного списка и выбор наиболее вероятных структур.

Реализация данного подхода имеет следующие особенности:

- возможность установления строения соединения как по отдельным видам спектров, так и по различным их комбинациям;
- использование наряду с БД таблиц спектро-структурных корреляций;
- решение задачи “шаг за шагом”, от определения молекулярной массы до построения и проверки структурных формул неизвестного соединения;
- получение на всех этапах ранжированных списков решений, среди которых с высокой степенью вероятности должны быть правильные ответы;
- максимальное сокращение списка структурных гипотез для последующей их проверки на соответствие экспериментальным спектрам;
- участие исследователя на всех этапах решения структурной задачи.

В рамках охарактеризованного выше подхода одну из основных функций выполняют поисковые процедуры. Они обеспечивают поиск соединений, спектры которых подобны

предъявленному. В ответ поисковой процедуры записываются спектры и структуры ограниченного числа соединений из БД (как правило, не более 20), получивших в результате сравнения спектров наивысшие оценки, отражающие степень их подобия (MF ≤ 100%). В зависимости от типа спектра (масс-, ИК, ПМР, ^{13}C ЯМР) и решаемой задачи, используются различные приемы сопоставления спектров, разные методы вычисления параметра MF и способы ранжирования результатов поиска [98,126,138,165-167]. Поисковый ответ представляет собой выборку, содержащую сведения о соединениях, наиболее близких в спектральном, а, следовательно и в структурном отношении к изучаемому объекту.

Логика анализа поискового ответа может быть проиллюстрирована следующим примером. На рис.2.1 приведен масс-спектр “неизвестного” соединения **C1** и сведения о первых пяти соединениях поискового ответа. Если допустить, что мы не знаем, что это за соединение, то из анализа поискового ответа можно сделать следующие предположения. Во-первых, молекулярная масса **C1** больше 115 ед., так как в масс-спектрах соединений поискового ответа отсутствуют пики молекулярных ионов (M^+) или их интенсивность очень мала. Во-вторых, весьма вероятно, что **C1 -** кислородсодержащее соединение, так как все пять соединений поискового ответа являются кислородсодержащими и, более того, они содержат по два атома кислорода. В третьих, можно допустить, что **C1** относится к производным 1,3-диоксолана, поскольку на первом и третьем местах ответа находятся соединения соответствующей структуры, и их спектры очень похожи на спектр **C1**. И, наконец, можно допустить наличие в молекуле **C1** изо-C_3H_7 или н-C_3H_7 групп, которые встречаются в структурах соединений поискового ответа. Последнее предположение подтверждается интенсивным пиком ионов с m/z = 43 а.е.м. Подобный ход рассуждений положен в основу всех рассматриваемых ниже методов решения задач по установлению строения неизвестных соединений.

рисунок дан в виде отдельного файла рис2-1

Рис. 2.1. **Первые пять соединений поискового ответа, полученного по масс-спектру “неизвестного” соединения C1.**

.

2.1.1. Определение молекулярной массы и молекулярной формулы соединения по масс-спектрометрическим данным

Пример, рассмотренный в предыдущем разделе, показывает, что компьютерное сопоставление спектров соединений БД с экспериментальным спектром позволяет сделать заключение о наличии или об отсутствии пика молекулярного иона в спектре анализируемого образца. Дальнейший анализ молекулярных масс и брутто-формул отобранных соединений дает сведения как о молекулярной массе изучаемого соединения (ММ), так и о возможном типе и количестве атомов элементов, образующих его молекулярную формулу (то есть о брутто-формуле). Очевидно, что эти сведения можно использовать в качестве граничных условий при генерации и ранжировании списка наиболее вероятных брутто-формул. Естественно, ММ соединения можно определить и традиционными экспериментальными средствами, а брутто-формулу – путем измерения точного значения массы молекулярного иона по масс-спектру высокого разрешения. Излагаемый ниже подход иной. Он опирается только на компьютерный анализ масс-спектра низкого разрешения исследуемого соединения.

В алгоритмах опознания пика молекулярных ионов (M^+), наряду с традиционными критериями - максимальным значением *m/z* наблюдаемых в моноизотопном спектре пиков и массами логичных (нелогичных) первичных потерь нейтральных фрагментов из M^+ - применяются и дополнительные критерии. Например: ожидаемая интенсивность пика $M^+(I_м{}^{+}*)$ и ожидаемая четность значения *m/z* (четное или нечетное) пика $M^+(E_M{}^{+})$. Эти параметры определяют путем компьютерного анализа поискового ответа.

Так, в работах [148-149] ожидаемую интенсивность пика M^+ оценивают по среднему значению интенсивностей пиков M^+ в спектрах трех первых (n=3) соединений поискового ответа, а ожидаемую четность *m/z* пика M^+ — по соотношению частот встречаемости в ответе (n = 6) соединений с четными ($n_{\text{чет}}$) и нечетными ($n_{\text{нечет}}$) величинами молекулярных масс. В качестве ионов-кандидатов в M^+ рассматривают все пики ионов в диапазоне *m/z* от m_{max} до m_{max} - 12, где m_{max} — максимальное значение *m/z* пика иона, наблюдаемое в скорректированном на природную распространенность изотопа ^{13}C спектре исследуемого соединения. Ранжирование кандидатов в M^+ осуществляют при помощи обобщенного параметра $R1=\alpha\beta\gamma\ (R_I + R_E)/2$, где α, β и γ-коэффициенты, принимающие значения 0 или 1. Коэффициент α=0, если иону-кандидату в исследуемом спектре соответствуют первичные нейтральные потери с массами 5÷12, 22÷25 а.е.м.; β=0 при $I_{M+} \leq 3\%$ от интенсивности

основного пика; γ=0, если ион-кандидат имеет четное значение *m/z*, а 6 первых соединений поискового ответа имеют нечетные значения молекулярных масс. Величины R_I и R_E— эмпирические параметры, используемые для количественной оценки достоверности предсказания I^x_{M+}- и E^x_{M+}, соответственно. Кандидаты в M^+, для которых $Rl \geq 10$, рассматриваются как вероятные. В этих случаях считается, что пик M^+ присутствует в спектре исследуемого соединения и опознан, т.е. молекулярная масса соединения определена. Чем больше значение параметра *R1*, тем больше доверие данному значению M^+. В случае, если ни один кандидат не имеет значение $Rl \leq 10$, считают, что пик M^+ отсутствует в спектре и требуется дополнительная процедура предсказания величины массового числа молекулярного иона.

Неустойчивость молекул некоторых веществ к ионизации электронным пучком приводит к полному распаду M^+ в ионизационной камере. В результате в масс-спектрах наблюдают только пики осколочных ионов, массы которых (*m*) отличаются от массы M^+ (M) на величины Δm, характеризующие массы нейтральных фрагментов, элиминируемых на первоначальных этапах распада M^+ по реакции:

$M \rightarrow m + \Delta m$.

Следовательно, для определения искомой массы иона M^+ необходимо знание величин Δm, суммирование которых с массами осколочных ионов позволяет получить возможные значения масс M^+. В работах [148-149] для генерации возможных кандидатов в M^+ используют один из наиболее интенсивных ($I \geq 0{,}3\%$) пиков осколочных ионов, наблюдаемый в моноизотопном спектре исследуемого соединения. Список масс нейтральных фрагментов, которые могут проявиться в спектре исследуемого соединения, определяют в результате анализа моноизотопных спектров 14 первых соединений поискового ответа: Одновременно для каждой $\Delta m_к$ рассчитывают и запоминают относительные частоты повторения данной потери (*p*) среди потерь *n* спектров соединений ответа ИПС. Список кандидатов в M^+ $\{M^+\} = \{M_i...,... M_s\}$ получают суммированием величин $\Delta m_к$ с m_{max} . Все *s* элементы множества M выступают в качестве кандидатов в молекулярные ионы. С целью ранжирования полученных кандидатов для каждого *j* элемента множества {M} формируется список $\{\Delta m^/ \}_j$, $\Delta m_l^/ =(M—m_l)_j$, где m_l — значение *m/z* *l*-пика в спектре изучаемого соединения, удовлетворяющее соотношению: $m_{max} \geq m_l \geq m_{max}/2$. Далее элементы списка $\{\Delta m \}$ последовательно сопоставляются с элементами списка $\{\Delta m'\}$,. Для совпавших по величинам первичных потерь ($\Delta m_к = \Delta m_l '$) рассчитывается суммарный весовой фактор вероятных потерь кандидата в молекулярные ионы:

$$WLY_j=\Sigma p^Y \log_2 \cdot P^Y,$$

где p^Y—относительные частоты встречаемости в *n*-спектрах поискового ответа совпавших первичных потерь и P^Y — относительные частоты встречаемости тех же потерь среди всех спектров БД. В случаях, если $\Delta m_к$ не равно Δm_l, вычисляется суммарный фактор

$$WLN_j = = \Sigma p^Y \log_2 \bullet P^Y ,$$

где $= p^Y$ и P^Y — те же параметры, что и ранее, но характеризующие невероятные для данного M_j первичные потери.

Окончательное ранжирование кандидатов M_j (j=1,..., s) осуществляют по величинам параметров $R2_j = (RL_j+RE_j)/2$, %, где $RL_j=WLY_j/(WLY_j+WLN_j)$. При этом, чем больше значение $R2_j$, тем выше доверие данному кандидату.

Эффективность этого подхода предсказания проверена на примере анализа масс-спектров 354 “неизвестных” соединений (267 спектров с пиком M^+, 87—без пика M^+), отобранных случайным образом из БД по масс-спектрометрии [148]. В этой выборке представлены соединения самых разнообразных химических классов, молекулярные массы которых лежали в диапазоне от 40 до 600 а.е.м. В соответствии с изложенным алгоритмом, для 166 соединений выборки выполнялась процедура идентификации пика M^+, а для 188— предсказание его положения в шкале массовых чисел анализируемых спектров

Таблица 2.1.

Вероятности определения молекулярных масс 354 “неизвестных” соединений по первым значениям ММ списка кандидатов в M^+

n	Контрольная выборка		P_{MM}
	спектры с пиком М+	спектры без пика М+	
1	94,4	52,9	88,2
3	99,6	70,1	95,2
5	99,6	78,2	96,4
10	99,6	86,2	97,6
{M}	100,0	97,7	99,7

Примечание: *n*– число рассматриваемых кандидатов из ранжированного списка, содержащего {M} кандидатов. Остальные обозначения – см. в тексте. “Вероятности” определены как отношение числа корректно решенных к общему числу решаемых задач.

В таблице 2.1. приведены результаты, полученные при анализе спектров контрольной выборки. Видно, что, если в спектре имеется пик M^+, то вероятность определения искомого значения молекулярной массы по первому кандидату ранжированного списка {M} составляет 94,4% и значительно выше, чем для спектров выборки, пики M^+ в которых отсутствовали (52,9%). Особо следует подчеркнуть тот факт, что для анализируемой группы

спектров в первом случае достигнута 100%-ная правильность определения ММ по полному списку {M}. Во втором — вероятность ошибочного определения составила 2,3% . На полученные результаты практически не повлияло, что 38% соединений, в спектрах которых присутствовали пики M^{+}, анализировались с помощью программы предсказания величины *m/z* для M^{+}. Позднее [149] анализ был проведен на существенно большей выборке, в частности, установлена возможность предсказания пика молекулярных ионов на более чем 5000 объектах, в масс-спектрах которых эти пики отсутствовали.

В четвертом столбце таблицы 2.1. приведены оцененные вероятности определения молекулярных масс (Рмм), которые можно ожидать при использовании описанного подхода в исследовательской практике. Они получены в предположении, что пики молекулярных ионов присутствуют в спектрах электронного удара 85% органических соединений. По данным этого столбца можно видеть, что разработанный метод позволяет определять истинное значение ММ более чем в 99% случаев, причем Рмм для первых трех ММ списка, предлагаемого компьютером, близка к 95%.

Существенно иной подход предложен в работе [150]. Его основное достоинство – ожидаемая меньшая зависимость результата определения молекулярной массы соединения от чистоты исследуемого соединения.

Теперь рассмотрим способы определения молекулярной формулы (МФ). Предсказание типов и количества атомов элементов, а также формальной ненасыщенности соединения получают из анализа брутто-формул (БФ) шести первых соединений поискового ответа, формируемого по заданному масс-спектру. Поиск в этом случае проводят с целью выявления из исходного множества химических элементов, например, {A°}= C, H, O, N, Si, S, P и F такого подмножества $\{A^{x}\}$, которое описывает состав молекулы исследуемого соединения. При этом необходимым условием для включения в список $\{A^{x}\}$ элемента из списка $\{A^{0}\}$ является наличие элемента по крайней мере в одной из БФ соединений ответа ИПС. Показано, что использование такого простого критерия позволяет предсказывать элементы подмножества $\{A^{x}\}$ с вероятностью, превышающей 90%. Из дополнительного анализа получают сведения о возможном числе атомов элементов списка $\{ A^{x} \}$, причем для каждого элемента этого списка (кроме водорода) предсказывают возможный интервал числа его атомов: (a_{max}, a_{min}), где a_{max} и a_{min} соответствуют максимальному и минимальному числу атомов данного элемента в БФ соединений ответа ИПС.

Анализ БФ поискового ответа позволяет также получить сведения о формальной ненасыщенности ($ФН^{x}$) исследуемого соединения, представляющей собой суммарное число эквивалентов двойных связей и колец:

ФН=a(4)+0,5-a(3)—0,5-a(1)+1,

где a(i) - число атомов элементов с валентностью i. При этом сведения о ФН, как и в случае Δa^x, определяют в виде возможного интервала значений:

$\Delta ФН_x = (ФН_{max}, ФН_{min})$,

где $ФН_{max}$, и $ФН_{min}$ —соответственно максимальное и минимальное значения ФН анализируемых соединений.

Для генерации брутто-формул исследуемого соединения используют следующие данные: ранжированный в соответствии с параметрами *R1* и *R2* список целочисленных значений масс молекулярных ионов {M}; списки возможных типов химических элементов $\{A^x\}$ и интервалов изменения числа их атомов $\{\Delta a^x\}$; интервал изменения суммарного числа эквивалентов двойных связей и колец ($\Delta ФН_x$). Генерация БФ обеспечивает получение исчерпывающего списка БФ - {БФ}, удовлетворяющих этим данным.

Ранжирование БФ осуществляется при помощи параметров *RF1* и *RF2* соответственно для случаев, когда список {M} получен в процессе идентификации M^+ и предсказания M^+. Параметры *RF1* и *RF2* — произведения *R1* и *R2* на фактор *Q*, характеризующий статистическую значимость отдельных химических элементов БФ в пределах шести первых соединений поискового ответа:

$Q = \nu_{фн}+\Sigma \nu_{ij}$,

где $\nu_{фн}$ - частота встречаемости величины ФН кандидата в БФ среди ФН БФ соединений ответа ИПС: ν_{ij} - соответствующая частота встречаемости *i*-го типа химического элемента БФ-кандидата с *j*-м числом атомов.

В случае, когда БФ-кандидат имеет состав C_cH_h, оценивается только частота встречаемости углеводородов ($\nu_{сн}$) и параметр *Q* вычисляется следующим образом: $Q = = \nu_{ФН}+\nu_{CH}$. Легко видеть, что чем больше значение параметра *RF1* или *RF2*, тем больше вероятность того, что данная БФ отвечает истинной брутто-формуле исследуемого соединения.

По результату анализа БФ соединений ответа ИПС с высокой вероятностью могут быть вынесены заключения о типе элементов $\{A^x\}$, диапазонах чисел атомов и ожидаемых величинах $\Delta ФН^x$. Например, вероятность правильного определения диапазона чисел атомов элементов C, O и N составляла 94,2; 91,0 и 97,7%, если анализируют спектры соединений, пики ионов M^+ в которых присутствуют. Соответствующие вероятности определения Δa^x по спектрам соединений, пики M^+ в которых отсутствуют, равны 90,4; 88,7 и 97%. Наряду с этим вероятность определения диапазона величины ФН, которому принадлежит ФН искомой брутто-формулы, в обоих случаях близка к 98%. Эти обстоятельства в сочетании с некоторой неоднозначностью в определении молекулярных масс, приводят к тому, что БФ исследуемого соединения при ее генерации включается в список БФ лишь в ~75% случаев.

В реальной исследовательской практике экспериментатор наряду с масс-спектром имеет дополнительную информацию об изучаемом соединении, полученную, например, из данных других спектральных методов или предыстории образца. Эти сведения могут способствовать однозначному выбору искомой БФ из списка, генерируемого ЭВМ (подобно тому, как это делается в масс-спектрометрии высокого разрешения).

Например, показано, что близкую к абсолютной вероятность получения искомой БФ в списке {БФ} можно достичь, если диапазон чисел атомов выявленных ЭВМ элементов увеличить на ± 2 единицы, а соответствующий диапазон ФН увеличить на ±1 единицу. В этом случае вероятность того, что искомая БФ окажется в списке {БФ}, сгенерированном ЭВМ, превышает 96%. Причем, на объектах контрольной выборки оказалось, что вероятность определения БФ “неизвестных”, в спектрах которых присутствует пик M^+, среди первых пяти БФ, равна 82,4%, а для соединений, в спектрах которых пик M^+ отсутствует — 59,8%. Оцененная на основе этих данных вероятность того, что при практическом использовании рассматриваемого подхода искомая БФ окажется в списке первых пяти БФ, близка к 80% .

Результат определения БФ по предъявленному спектру в предположении, что для исследуемого соединения имеют сведения об ожидаемом значении ММ с погрешностью ±5% (эти данные с более высокой точностью могут быть получены, например, традиционными аналитическими методами эбулиоскопии, криоскопии, осмометрии или эффузиометрии), заметно изменяется лишь для соединений, в спектрах которых пики M^+ отсутствуют. В этом случае вероятность определения искомых БФ соединений контрольной выборки среди первых, первых трех или пяти БФ списка возрастает на ~20%. В то же время относительно малая доля таких соединений слабо влияет на рост оцененной вероятности определения БФ, которую можно достичь в реальной исследовательской практике

Проиллюстрируем возможность использования описанного выше метода в предположении, что экспериментатору известна некоторая информация об элементном составе изучаемого соединения. Заметим при этом, что многообразие возможных ситуаций, предшествующих обращению к методу определения БФ, невозможно охватить в полном объеме, поэтому представленные в таблице 2.2. результаты компьтерных экспериментов приведены при следующих простых допущениях. Первое — по данным способа синтеза известно, что соединение содержит определенное число атомов одного из элементов N, O или C. Например, информация о числе атомов C может быть получена по данным спектроскопии углеродного магнитного резонанса. Второе — проведен CHN-анализ исследуемого соединения (например, с использованием автоматического CHN-анализатора).

Таблица 2.2.

Вероятности определения искомой БФ среди *n* первых брутто-формул списка {БФ} при наличии априорной информации об элементах БФ

n	$P^*_{БФ}$			
	N	**O**	**C**	**CHN**
1	65	77	78	86
3	79	87	91	94
5	86	90	94	96
10	90	93	95	97
{БФ}	97	97	97	98

Данные этой таблицы показывают, что при использовании априорной информации о соединении наблюдается заметный рост вероятности определения БФ за счет удаления из списка брутто-формул, не удовлетворяющих дополнительным требованиям. Оцененная по спектрам выборки вероятность определения искомой МФ ($P^*_{БФ}$) среди брутто-формул списка возрастает и в случае привлечения данных CHN-анализа составляет 94% (при n=3).

В этой же связи отметим метод установления МФ соединения по масс-спектру низкого разрешения и спектрам ЯМР (^{1}H, ^{13}C), предложенный в работах [168-169]. Он использует информацию, содержащуюся в спектрах ЯМР (число и мультиплетности сигналов внерезонансного спектра ^{13}C, интегральные интенсивности сигналов спектра ^{1}H), как для корректировки числа некоторых атомов перед генерацией возможных брутто-формул, так и для проверки генерированных МФ на непротиворечивость этим спектрам [168]. Эффективность соответствующего программного обеспечения проверена на многочисленных примерах разнообразных органических веществ, взятых из практики и из книги [171]. Ниже приведены относительные частоты появления искомых молекулярных формул ($P_{МФ}$, %) среди n первых в ранжированном списке решений для двух серий экспериментов (МС - молекулярная формула определяется только по масс-спектру;

МС+ЯМР - с привлечением спектров ЯМР):

n	1	3	5	10	>10	
$P_{МФ}$	46	69	71	78	94	(МС)
$P_{МФ}$	83	89	94	96	97	(МС+ЯМР)

Очевидно, что во втором случае результаты лучше. Так, для подавляющего большинства решенных задач (~94%) определяемая МФ оказывается среди первых пяти кандидатов, выдаваемых ЭВМ. Важно отметить и то, что в данном эксперименте в половине всех рассмотренных случаев получено единственное и корректное решение. Списки возможных МФ включали в среднем около МФ. Это вполне удовлетворительный результат, поскольку он получен в условиях, когда сведения о молекулярной массе соединения не задаются, а определяются (и напомним – не всегда однозначно) с помощью компьютера.

Определение МФ соединения, как видим, обеспечивает такой уровень достоверности результата, который позволяет использовать эти данные не только при идентификации соединения с помощью ИПС [167], но и на последующих этапах решения структурной задачи в случаях, когда соединение не идентифицировано.

2.1.2. Определение "фрагментарной" формулы соединения и дескрипторов формальной ненасыщенности

При решении структурных задач с помощью ЭВМ в качестве базовой характеристики выступает, как правило, молекулярная формула, на основе которой выполняются все операции по установлению строения соединения, в том числе и построение его возможных структур. Однако МФ практически не несет информации о строении, указывая только типы и число атомов, из которых состоит молекула анализируемого вещества. Поэтому казалось целесообразным найти более информативную в структурном отношении характеристику, которая, как и МФ, могла быть определена по спектральным данным. Исследования показали, что в качестве таковой может выступать так называемая фрагментарная формула (ФФ) - набор структурных единиц (дескрипторов), полностью согласованных по элементному составу с МФ соединения [172]. Дескрипторами при этом являются атомы элементов (кроме водорода) и микрофрагменты (CH_3, CH_2, CH, CO, CHO, CN, SO_2, NO_2 и т.п.), используемые нами при описании химических графов. В этом случае, например, для стирола (МФ = C_8H_8) ФФ=$(CH_2)_1(CH)_6(C)_1$. Легко видеть, что фрагментарная формула по сравнению с молекулярной более информативна в структурном отношении. Это находит отражение, в частности, в числе генерируемых на ее основе структурных изомеров. Так, для приведенного выше примера с использованием сведений о МФ можно построить 7437 изомеров, в то время как на основе ФФ - только 280.

В работе [172] предложена еще одна характеристика соединения - брутто-формула формальной ненасыщенности (БФФН). Она представляет собой набор дескрипторов формальной ненасыщенности (ФН), полностью описывающих значение ФН соединения. Последняя легко определяется по молекулярной формуле: ФН = q(4)+0.5q(3)-0.5q(1)+1, где q(i) - число атомов химических элементов с валентностями i (в нашем случае i=4 (C, Si), i=3 (N, P), i=1 (H, Cl, Br, F, I). В качестве дескрипторов ФН используются: двойная связь (ДС), тройная связь (ТС), насыщенный цикл произвольной размерности (ЦП) и шестичленное углеродсодержащее ароматическое кольцо (АК), имеющие значения ФН 1, 2, 1 и 4 соответственно. Например, для стирола (МФ = C_8H_8) ФН = 5, а БФФН = $(ДС)_1(АК)_1$. Легко убедиться, что в данном случае с использованием сведений о ФФ (см. выше) и БФФН можно построить только одну структурную формулу.

Алгоритм определения рассмотренных выше характеристик по масс-спектрам низкого разрешения описан в работе [172]. Его некоторые особенности рассмотрим на примере определения фрагментарной формулы.

В качестве исходных данных для построения возможных ФФ изучаемого соединения используется множество G = {D, l, h}, где D_i – тип структурного дескриптора, а l_i и h_i - минимально и максимально допустимые числа дескрипторов данного типа. На многочисленных примерах показано, что условием включения D_i во множество G (наряду с непротиворечием МФ исследуемого соединения) может быть его появление, по крайней мере, один раз в ФФ десяти первых соединений поискового ответа. Значения параметра h_i при этом определяются максимальным числом дескрипторов i типа в ФФ соединений соответствующего ответа. Фрагментарные формулы, построенные путем перебора различных сочетаний элементов множества G, ранжируются по значениям параметра Rff, определяемого в результате сравнения каждой построенной ФФ с соответствующими данными k-первых соединений поискового ответа. Наибольший вклад в Rff дают такие элементы ФФ, которые наиболее часто встречаются в ФФ соединений поискового ответа, а спектры последних наилучшим образом совпадают со спектром исследуемого соединения. При этом мы полагаем, что фрагментарные формулы, имеющие наивысшие значения Rff , могут выступать в качестве наиболее вероятных решений.

Эффективность рассмотренного метода оценена на примерах предсказания ФФ и БФФН ~700 “неизвестных” соединений, отобранных случайным образом из базы масс-спектрометрических данных. Экспериментально показано, что искомые ФФ и БФФН обнаруживаются в формируемых списках возможных решений в 80 и 94% случаев соответственно. При этом относительные частоты появления правильных ответов ($P_{ФФ}$ и $P_{БФФН}$) среди n-первых в ранжированных списках составляют:

n	1	3	5	10	
$P_{ФФ}$	52%	75%	86%	94%	(МС)
$P_{БФФН}$	75%	95.5%	-	-	(МС)

Результативность определения БФФН существенно лучше и достигает практически придельных значений при рассмотрении не более трех первых гипотез. Для определения ФФ требуется рассмотрение не менее десяти соответствующих решений. Эти результаты вполне удовлетворительны, т.к. они получены только на основе молекулярной формулы соединения и его масс-спектра. Понятно, что привлечение дополнительных спектральных данных позволит повысить результативность метода. Особенно эффективно использование спектров ^{1}H и ^{13}C ЯМР. В этом случае могут быть получены надежные сведения о типах и числе некоторых микрофрагментов (CH_3, CH_2, CH, C, CHO, CN и CO), идентификация которых по масс-спектрометрическим данным часто затруднена. Ниже приведены результаты тестовых испытаний разработанного метода определения фрагментарной формулы соединения в условиях, когда исследователь располагает масс- и ЯМР-спектрами [173,174]:

n	1	3	5	
$P_{ФФ}$	68%	95%	99%	(МС+ЯМР)
$P_{МФ}$	88%	95%	99%	(МС+ЯМР)

Близость соответствующих результатов при n = 3 и 5, на наш взгляд, однозначно характеризует разработанный метод определения фрагментарной формулы соединения по спектрам молекул. Важно подчеркнуть, что в процессе определения ФФ автоматически уточняются и результаты предсказания МФ, что сопровождается сокращением первоначальных списков решений. Это приводит к более высоким величинам $P_{МФ}$ и позволяет рекомендовать данный метод и для определения молекулярной формулы соединения. В качестве примера ниже приведены результаты определения ФФ и МФ “неизвестного” соединения **C1** (см. рис.2.1) по следующим спектральным данным [26]:

масс-спектр (m/z-интенсивность):

27-110 28-120 39-60 41-170 43-650 45-50 55-55 71-180 99-40
113-30 115-1000 116-120;

^{1}H ЯМР спектр (химический сдвиг/интеграл):

0.75-0.95/30.0 1.85-2.15/5.0 3.87-4.07/10.0;

^{13}C ЯМР спектр (химический сдвиг/мультиплетность/интенсивность):

17.5/к/60.0 35.0/д/29.0 67.8/т/48.0 117.0/с/9.0

ФФ-кандидатов: 2	МФ	ММ
$(CH_3)_4(CH_2)_2(CH)_2(C)_1(O)_2$	$C_9H_{18}O_2$	158
$(CH_3)_4(CH_2)_2(CH)_2(C)_1(O)_1$	$C_9H_{18}O_1$	142

Можно видеть, что несмотря на отсутствие в масс-спектре пика молекулярного иона (m/z = 158), компьютер успешно справился с поставленной задачей и выдал исследователю только два решения, первое из которых содержит ФФ и МФ, на самом деле отвечающие “неизвестному” соединению.

Следует отметить, что описанные выше методы могут использоваться при решении целого ряда задач аналитической практики, когда конечной целью является не установление строения соединения, а получение отдельных сведений, отражающих, например, химический класс анализируемого вещества; наличие определенного типа и числа функциональных групп (OH, CO, NO_2, SO_2 и т.п.), двойных, тройных связей и циклов.

2.1.3. Определение крупных структурных фрагментов соединения

В системах “искусственного интеллекта” структурная информация извлекается, как правило, с помощью корреляционных таблиц, содержащих сведения о спектральном поведении сравнительно небольших (по числу скелетобразующих атомов) фрагментов, типа CH_3O-, CH_3-CO-, CH_2=CH-, -CH_2-O-CH_2- и т.п.. Понятно, что использование подобных фрагментов для построения возможных структур изучаемого соединения требует, в общем случае, существенных затрат машинного времени и приводит к большому списку гипотез, последующий анализ которых крайне затруднителен. Отмеченные недостатки можно устранить, обеспечив опознание достаточно крупных связных фрагментов соединения. И здесь просматриваются два пути. Первый состоит в применении методов “распознавания образов” [175-176] или еще лучше “искусственных нейронных сетей” [177-180], позволяющих по результатам предварительного обучения с высокой степенью достоверности опознавать фрагменты из предварительно сформированного списка. Второй путь связан с использованием банков данных (БД), содержащих информацию о спектрах и структурах большого числа ранее описанных соединений [181-185]. Сведения о строении анализируемых объектов в этом случае определяются из анализа структур соединений БД, подобных исследуемому по тем или иным спектральным признакам. Такой подход не имеет принципиальных ограничений на типы и размеры выявляемых фрагментов, а результативность его определяется наличием в БД структурных аналогов изучаемого соединения, методами их отбора и анализа результатов поиска.

2.1.3.1. Извлечение структурной информации из ИК- и масс-спектров

Предназначенные для решения поставленной задачи поисковые процедуры по ИК- и масс-спектрам (далее, ПОИСК-ИК и ПОИСК-МС) должны обеспечивать отбор из БД таких

соединений, последующий анализ которых позволял бы получать наиболее полные и достоверные сведения о строении изучаемых соединений. В случае ИК спектроскопии это достигается путем использования банков данных, содержащих практически полную информацию о спектральных кривых в координатах $\nu(см^{-1})/T$ (%), и трехкомпонентного фактора для оценки степени подобия сравниваемых спектров X и R [126]. Каждый компонент несколько по-разному оценивает степень совпадения спектров: K1 – только по положениям полос поглощения; K2 – только по положениям полос с близкими значениями интенсивностей (в наших экспериментах T ± 30%); K3 - по положениям и интенсивностям всех сравниваемых полос. Ранжирование результатов поиска с помощью параметра $MFI = (\max\{K1, K2, K3\} + \overline{K})/2$ позволяет отбирать в поисковые ответы соединения по максимальному значению одного из трех компонентов; другие компоненты при этом учитываются средним значением $\overline{K} = (K1 + K2 + K3)/3$. Как показали эксперименты, такой прием обеспечивает появление в ответе поисковой процедуры структурных аналогов изучаемого соединения в разнообразных ситуациях, связанных с чистотой анализируемого образца и условиями регистрации спектра.

При решении задач по масс-спектрам, наряду с различными стратегиями поиска (прямой, комбинированный, обратный), предусмотрена возможность использования двух групп спектральных признаков, связанных с абсолютными (значения m/z) и относительными (Δm=MM–m/z) положениями линий в спектре. Эти признаки позволяют получать дополняющие и уточняющие друг друга сведения о строении анализируемых объектов [186,187]. При этом совместное использование указанных признаков, обеспечивает, как правило, отбор самых близких структурных аналогов изучаемого соединения. Однако в общем случае имеет смысл рассматривать комбинированные поисковые ответы, составленные из “лучших” соединений, отобранных по различным видам признаков с использованием различных стратегий поиска, поскольку в каждом конкретном случае трудно предугадать какие признаки, и какая стратегия поиска приведут к наилучшим результатам.

В настоящее время для анализа поисковых ответов по масс- и ИК-спектрам наиболее развиты три подхода. Первый связан с идентификацией фрагментов заданного строения из предварительно сформированного списка [181, 184], второй – с определением фрагментов, используемых при описании структур соединений, хранящихся в банках данных [183], а третий – фрагментов произвольного строения, выявляемых путем “пересечения” структур соединений поискового ответа [125,188,189]. Последний подход, несмотря на трудоемкость, нам кажется наиболее перспективным. Продемонстрируем его возможности на примере решения структурных задач по ИК- и масс-спектрам [167].

Во всех экспериментах анализ поисковых ответов осуществлялся с помощью

высокоэффективной программы (Кохов В.А., 1989), обеспечивающей выделение из пары структур исчерпывающего списка максимально общих (по числу вершин) фрагментов. При этом для последующего рассмотрения остаются только связные фрагменты, составленные не менее чем из пяти вершин. Такое ограничение кажется разумным, поскольку фрагменты меньших размеров могут быть определены с помощью корреляционных таблиц. Исследования показали, что наиболее полные и достоверные сведения о строении изучаемых объектов выявляются, как правило, из анализа ограниченного числа k-первых структур соединений поисковых ответов, имеющих высшие значения факторов спектрального (в наших экспериментах: $k \leq 10$; $MFI \geq 50\%$ и $MFM \geq 30\%$ для процедур ПОИСК-ИК и ПОИСК-МС соответственно). Полученные в результате $2k*(2k-1)/2$ попарных пересечений $2*k$ структур неизоморфные фрагменты проверяются на соответствие молекулярной формуле и формальной ненасыщенности. В список фрагментов-кандидатов заносятся только такие, которые встречаются, по крайней мере, один раз в структурах соединений обоих поисковых ответов.

Задача выявления наиболее вероятных для изучаемого соединения фрагментов решается путем ранжирования их по значениям следующего параметра:

$$RFr = (N_I + N_M + N_C) * Q_F,$$

где N_I и N_M – число соединений в ответах ПОИСК-ИК и ПОИСК-МС, содержащих данный фрагмент соответственно; $N_C = \min(N_I, N_M)$; Q_F - "размер" фрагмента, оцененный числом всех, кроме водорода, составляющих его атомов ($Q_F \geq 5$). По значению параметра RFr предпочтение отдается наиболее крупным и часто встречающимся в структурах соединений ответов ПОИСК-ИК и –МС фрагментам. Чем выше значение этого параметра, тем больше вероятность того, что данный фрагмент принадлежит структуре изучаемого соединения.

Эффективность рассмотренного подхода проверена на примерах решения 50 контрольных задач. В качестве “неизвестных” выступали разнообразные органические соединения с обобщенной МФ следующего вида:

$$C_{4-20}\ H_{5-32}\ O_{0-7}\ N_{0-2}\ P_{0-1}\ S_{0-1}\ Br_{0-1}\ Cl_{0-5}\ F_{0-3}\ Si_{0-1}.$$

Оцененные числом всех (кроме водорода) атомов элементов МФ “размеры” структур неизвестных соединений (Q_S) изменялись в диапазоне от 8 до 25 ед. и составляли в среднем 13 атомов. Анализ полученных при решении контрольных задач фрагментов показал, что половина из них оказалась некорректными. Таким образом, факт появления в структурах соединений поисковых ответов по ИК- и масс-спектрам общего и достаточно крупного фрагмента еще не гарантирует его присутствия в молекуле исследуемого соединения. Ситуация улучшается, если рассматривать более часто встречающиеся фрагменты. Например, в трех сериях экспериментов (**I** – $N_C \geq 1$, N_I или $N_M \geq 3$; **II** – $N_C \geq 2$; **III** – $N_C \geq 3$)

получено:

	P_1	P_3	P_5	W_5	P_0
I	79	85	91	52	68
II	82	85	96	47	54
III	***86***	***85***	***100***	***46***	***44***

где, P_n - процент правильных решений от числа решенных задач (под правильным решением понимается наличие, по крайней мере, одного корректного фрагмента среди n-первых ранжированного списка); P_0 - процент правильных решений от общего числа задач; W_n - средний размер корректных фрагментов, выявляемых при рассмотрении n-первых фрагментов ранжированного списка ($W_i = 100*Q_F/Q_S$).

Приведенные данные показывают, что с увеличением N_C, N_I и N_M повышаются величины P_n, достигая наивысших значений для фрагментов **III** категории. При этом, однако, уменьшаются размеры опознаваемых фрагментов, сокращается число решаемых задач и остаются, к сожалению, случаи выявления некорректных фрагментов. Показано [167], что избежать подобных результатов не удается и при использовании статистических критериев, учитывающих частоты встречаемости фрагментов в структурах соединений соответствующих БД [190]. Поэтому при решении практических задач можно рассматривать фрагменты всех трех категорий, понимая, эта информация носит предположительный характер и может быть использована при генерировании структурных гипотез только в виде списка альтернативных фрагментов. В качестве примера в таблице 2.3 приведены лучшие по значениям параметра RFr корректные и некорректные фрагменты, предсказанные для шести “неизвестных” соединений из совместного анализа ответов процедур ПОИСК-ИК и ПОИСК-МС.

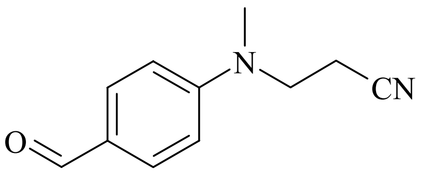
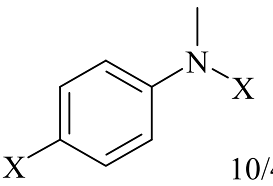
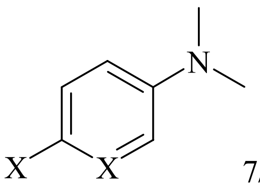
Таким образом, при выявлении фрагментов, подтвержденных масс- и ИК-спектрами, статистический подход оказывается недостаточным для достижения желаемых результатов (см., например, соединение **C2** в таблице 2.3). В связи с этим опишем кратко возможность использования спектральных данных для оценки достоверности выявляемых фрагментов на примере решения задач по ИК спектрам [126].

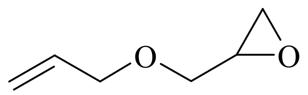
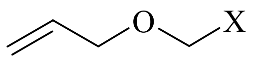
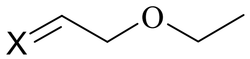
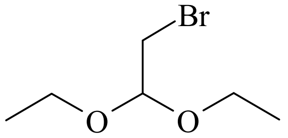
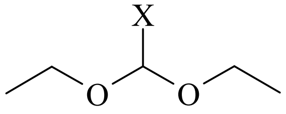
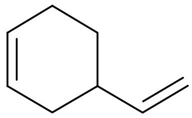
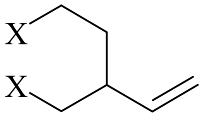
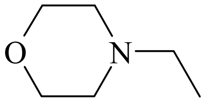
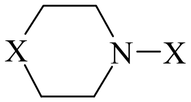
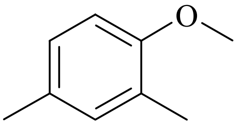
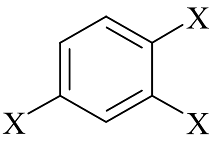
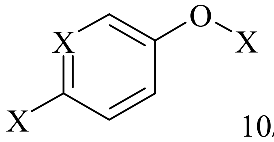
На первом этапе в спектрах соединений поискового ответа, содержащих проверяемый фрагмент ($Q_F \geq 5$, $N_I \geq 3$), отмечаются общие для всех спектров признаки - близкие по значениям частот поглощения полосы. Близкими считаются полосы, попадающие в интервалы частот шириной 60 и 30 $см^{-1}$ в диапазонах 4000÷2000 и 2000÷400 $см^{-1}$ соответственно. Затем выявленные наборы признаков (далее, спектральный отклик фрагмента) сравниваются со спектром X и ранжируются по значениям параметра MFI, характеризующего степень их “вложимости” в спектр X. Разумно допустить: чем лучше

спектральный отклик вкладывается в спектр исследуемого соединения, тем больше вероятность того, что соответствующий фрагмент принадлежит его структуре. В качестве примера на рис.2.2 приведены результаты работы обсуждаемой процедуры при установлении строения соединения **C8**. В данном случае из структур соединений поискового ответа выделено два достаточно крупных фрагмента, спектральные отклики которых не противоречат известным корреляциям ИК спектроскопии. Так, присутствие паразамещенного бензольного кольца во фрагментах **1** и **2** подтверждается полосами поглощения в интервале частот 1625÷1595, 1605÷1575, 1525÷1495 и 835÷805 $см^{-1}$. В спектральных откликах этих фрагментов им соответствуют полосы 1607, 1511 и 813 $см^{-1}$. Наличие группы **OH** во фрагменте **1** также не противоречит предъявленному спектру, поскольку в нем имеется полоса поглощения в интервале 3400÷3340 $см^{-1}$. О присутствии группы **CH_2** во фрагменте **2** может свидетельствовать, например, полоса в интервале 2930÷2870 $см^{-1}$. На основе выявленных фрагментов и молекулярной формулы не представляет труда построить структурную формулу соединения, так как группа –**CN** легко идентифицируется иследователем по очень интенсивной полосе в интервале 2270÷2210 $см^{-1}$.

Таблица 2.3.

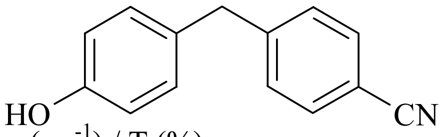
Примеры “лучших” корректных и некорректных фрагментов, выявленных из совместного анализа ответов процедур ПОИСК-МС и ПОИСК-ИК

	Структурная формула “неизвестного” соединения	**Предсказанные фрагменты**	
		Корректные	**некорректные**
C2		 10/4[a]	 7/4

		 3/2	 1/2
		 4/11	——
C5		 1/3	——
		 5/2	——
		 10/4	 10/1

Примечание: a - частотные характеристики фрагмента N_I/N_M (см. в тексте); X – любой, отличный от атома водорода, заместитель.

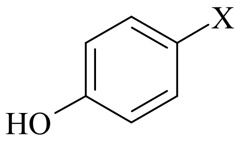
"Неизвестное" соединение:

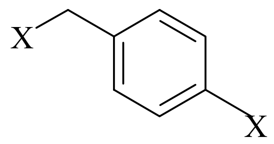


(C8)

ИК-спектр для поиска: ν_{min} - ν_{max} (см$^{-1}$) / T (%)

3400-3340/12	2930-2870/55	2270-2210/15	1625-1595/39	1605-1575/47
1525-1495/20	1455-1425/40	1425-1395/53	1365-1335/60	1275-1245/49
1225-1195/32	1195-1165/41	1125-1095/52	1025-995/67	885-855/56
835-805/29	785-735/55	755-725/52	645-615/56	585-555/32
545-515/40	515-485/62	465-435/70		

Фрагменты-кандидаты	MFI,%	Спектральные отклики фрагментов			
		3375/17	1607/61	1511/7	1450/40
		1362/60	1293/63	1266/63	1180/52
		1101/43	957/73	813/27	645/35
		500/64			



3026/32	2975/33	1682/1	1607/61
1511/7	1450/40	1295/63	1266/19
1225/33	1180/52	1172/39	1101/53
1023/81	884/53	813/27	776/39
645/35	607/74	500/64	

Рис. 2.2. Результаты анализа поискового ответа, полученного по ИК-спектру "неизвестного" соединения С8

Этот и другие примеры убеждают нас, что использование спектральных откликов может приводить к более обоснованным решениям, чем в случаях, когда достоверность фрагментов оценивается только по частотам их встречаемости в структурах соединений поискового ответа. Следует отметить, что выявляемые на сравнительно небольших выборках соответствия "фрагмент - спектральный отклик" носят предварительных характер и для дальнейшего их использования в качестве компонентов корреляционных таблиц требуется дополнительная проверка на всей базе данных или экспертная оценка специалиста. Рассмотренный выше прием применим и к анализу результатов работы поисковых процедур по масс-спектрам [191,192].

Аналитические характеристики обсуждаемого метода, полученные при решении упомянутых выше задач по ИК- и масс-спектрам соответственно, приведены ниже:

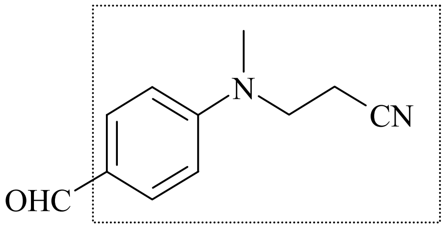
P_1	P_3	P_5	W_5	P_0	
68	85	87	54	62	ИК
62	78	80	62	78	МС

Неудивительно, что значения параметров P_n уступают аналогичным данным, полученным при совместном анализе ответов процедур ПОИСК-ИК и ПОИСК-МС. Однако требование наличия в ответах обеих процедур однотипных структурных аналогов исследуемого соединения является достаточно жестким и может приводить к выявлению сравнительно небольших фрагментов. В этом отношении предпочтительнее варианты анализа поисковых ответов по отдельности, так как в этом случае больше шансов получить сведения о крупных фрагментах изучаемого соединения. В качестве примера ниже приведены структурные формулы шести "неизвестных" соединений, в которых обведены фрагменты, выявленные по результатам работы отдельных поисковых процедур ПОИСК-ИК и ПОИСК-МС.

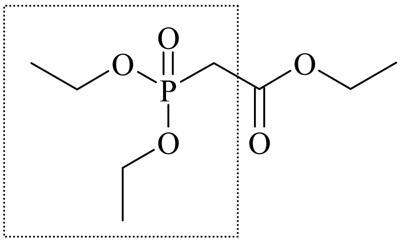
Не следует забывать, что в рамках рассмотренных подходов наряду с корректными выявляются и некорректные фрагменты. Большинство из них не противоречит известным спектро-структурным корреляциям и может быть отвергнуто только специалистом при тщательном анализе спектров. Другой путь - привлечение дополнительных данных. Так,

например, по мультиплетностям сигналов спектра ^{13}C ЯМР легко отбросить предсказанный для соединения **С3** некорректный фрагмент (см. табл.2.3). Однако более эффективным “фильтром” при проверке выявляемой структурной информации может быть фрагментарная формула соединения, надежно определяемая по его масс- и ЯМР-спектрам .

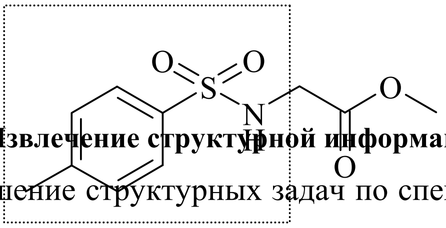
ПОИСК-ИК



(С2)

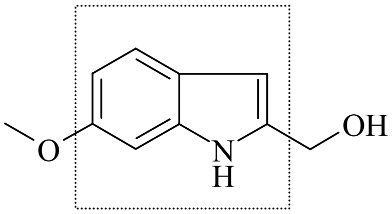


(С9)

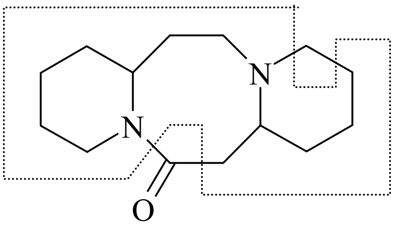


(С10)

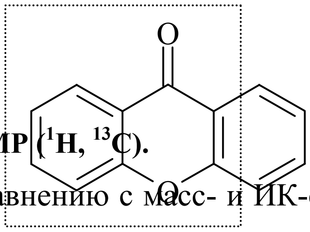
ПОИСК-МС



(С11)



(С12)



(С13)

2.1.3.2. Извлечение структурной информации из спектров ЯМР (^{1}H, ^{13}C).

Решение структурных задач по спектрам ЯМР (по сравнению с масс- и ИК-спектрами) выглядит намного проще, поскольку в используемых базах данных [68] имеется информация об “отнесениях” резонирующих атомов к отдельным сигналам, описанным значениями химических сдвигов в шкале м.д. (δ), мультиплетностями (М = с - синглет, д - дублет, т - триплет, к - квартет, м - мультиплет) и интенсивностями (In = число эквивалентных атомов). Следовательно, обеспечив отбор из базы данных спектров подобных предъявленному, в структурах соответствующих соединений можно пометить узлы (атомы, микрофрагменты), отвечающие совпадающим в спектрах X и R сигналам, и подсказать тем самым путь выделения требуемых фрагментов, удовлетворяющих спектру ЯМР изучаемого соединения. Для получения наиболее полной и достоверной структурной информации в процессе сопоставления спектров и при ранжировании результатов поиска по спектрам ^{1}H и ^{13}C ЯМР-спектрам (далее, ПОИСК-1Н и ПОИСК-13С) необходимо учитывать все спектральные параметры (δ, M, In) [166, 193].

Разработанный нами алгоритм [193] ориентирован на выделение из структур соединений поискового ответа связных фрагментов, составленных из “меченых” и “неактивных” узлов, отстоящих от “меченых” не далее, чем на одну связь. В качестве

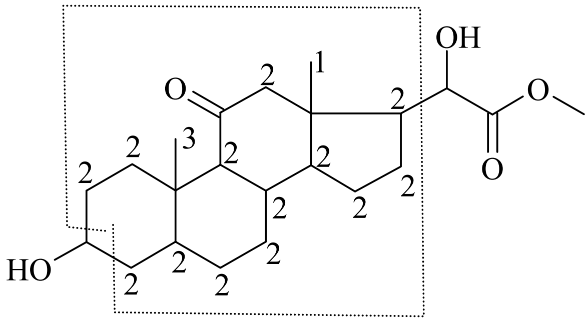
"неактивных" для спектроскопии ПМР выступают узлы, не содержащие атомов водорода, а для ^{13}C ЯМР - гетероатомные узлы (N, O, S, NO_2 и т.п.). При этом учитываются также сведения о молекулярной формуле изучаемого соединения, полагая, что эта информация может быть получена по спектральным данным с помощью ЭВМ (см. выше). В результате анализа поискового ответа формируется список неизоморфных фрагментов, который затем ранжируется. Преимущество отдается более крупным фрагментам, спектральное поведение которых наиболее адекватно отражено в спектре изучаемого соединения. Выдаваемые исследователю фрагменты сопровождаются информацией о номерах сигналов спектра X, к которым отнесены составляющие их узлы.

Эффективность соответствующего программного обеспечения проверена на многочисленных примерах решения задач с помощью поисковых процедур по спектрам ^1H- и ^{13}C-ЯМР. В качестве "неизвестных" выступали разнообразные органические соединения, "размеры" структур которых отражают следующие данные: $Q_S = 7 \div 39$, $\overline{Q}_S = 16$ (^1H ЯМР) и $Q_S = 9 \div 26$, $\overline{Q}_S = 13$ (^{13}C ЯМР). Ниже приведены значения параметров P_n, полученные при решении 50 задач по спектрам ^1H ЯМР [193]:

$P_1 = 44\%$, $P_3 = 70\%$, $P_5 = 74\%$, $P_{10} = 96\%$. (^1H ЯМР)

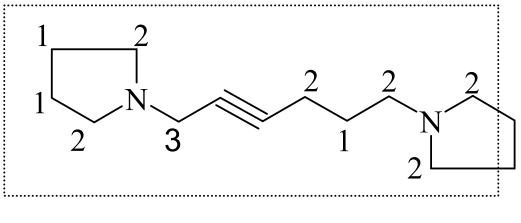
Видно, что среди первых 10 фрагментов ранжированного списка практически всегда (96% случаев) присутствует, по крайней мере, один, принадлежащий структуре "неизвестного" соединения. В зависимости от числа рассматриваемых фрагментов n = 1, 3, 5 и 10 размеры корректных несколько уменьшаются и составляют в среднем 68, 61, 60 и 55% от структуры изучаемого соединения. Ниже в качестве примера приведены спектры ПМР (δ_{min}- δ_{max}/M/In) и структурные формулы двух "неизвестных", в которых обведены фрагменты, выявленные из анализа соответствующих поисковых ответов.

1. 0.70-0.70/c/3
2. 0.90-2.70/м/22
3. 1.18-1.18/c/3
4. 2.87-2.87/м/1
5. 3.62-3.62/м.1
6. 7.78-3.78/c/3
7. 4.09-4.09/м/1



(C14)

1. 1.34-19.8/м/10
2. 2.04-2.37/м/12
3. 3.28-3.28/м/2



(C15)

Отметим, что результативность решения задач по спектрам ^{1}H ЯМР зависит от формальной ненасыщенности соединений:

	P_1	P_3	P_5	P_{10}	
ФН $\leq$ 4	50	85	90	100	(30 соединений)
ФН $\geq$ 5	40	48	48	80	(20 соединений)

Видно, что при анализе спектров соединений с относительно большим содержанием атомов водорода (ФН ≤4) величины P_n превышают аналогичные для всей выборки (см. выше), в то время как для соединений со значениями ФН ≥ 5 они существенно хуже. Последнюю выборку составляли, в основном, соединения ароматического ряда, структуры которых включают большое число "неактивных" для спектроскопии ПМР узлов (C, CO, N). С другой стороны, если алгоритм ориентировать на выделение фрагментов, составленных только из "меченых" узлов, то результат, как правило, тривиален и может быть получен традиционными способами.

Отмеченные выше недостатки практически устраняются при решении задач по спектрам ^{13}C ЯМР. Здесь достигается более высокий уровень достоверности структурных предсказаний независимо от ФН анализируемых веществ, о чем свидетельствуют значения параметров P_n:

$$P_1 = 60\%, \quad P_3 = 89\%, \quad P_5 = 90\%. \qquad (^{13}\text{C ЯМР})$$

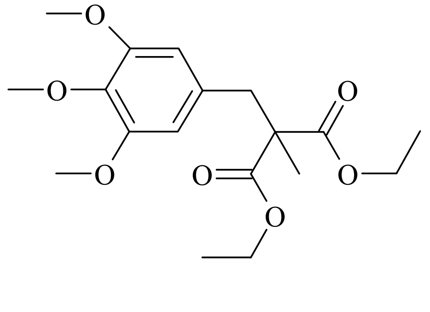
Размеры выявляемых корректных фрагментов также высоки и составляют в среднем 65, 58 и 58% от молекулы "неизвестного" соединения при рассмотрении 1, 3 и 5 первых фрагментов ранжированного списка. Однако для установления строения больших молекул даже подобных сведений недостаточно. Дополнительные данные могут быть получены, в частности, путем повторного поиска по спектру ^{13}C ЯМР, после исключения из него сигналов "отнесенных" к фрагменту, выявленному в первоначальном поиске. В результате двух поисков выявляются подтвержденные различными сигналами спектра X фрагменты (см. табл.2.4), что позволяет использовать их одновременно при построении возможных структур соединения.

Таблица 2.4.

Примеры решения двух задач с помощью процедуры ПОИСК-13С

Спектр ^{13}C ЯМР — Структурная формула "неизвестного" соединения

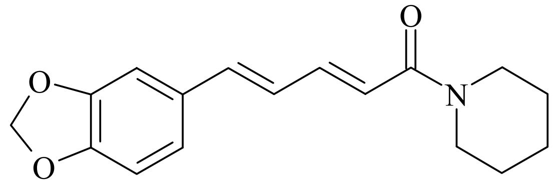
1 - 13.8/к/2	7 - 61.0/т/2
2 - 19.6/к/1	8 - 107.3/д/2
3 - 41.3/т/1	9 - 131.7/с/1
4 - 54.7/к/1	10- 137.0/с/1
5 - 55.8/к/2	11- 152.7/с/2



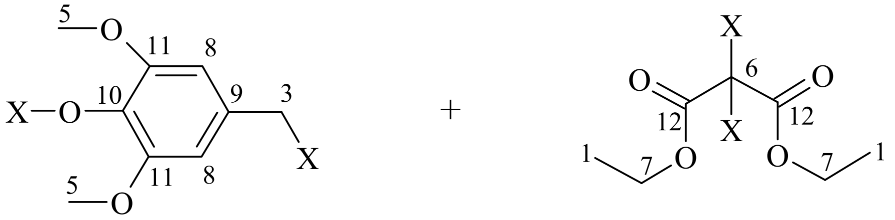
6 - 60.4/с/1 12- 171.6/с/2

1 - 24.6/т/1 8 - 122.4/д/1
2 - 26.2/т/2 9 - 125.4/д/1
3 - 44.8/т/2 10-131.0/с/1
4 -101.2/т/1 11-138.0/д/1
5 -105.6/д/1 12-142.3/д/2
6 -108.4/д/1 13-148.1/с/2
7- 120.2/д/1 14-165.2/с/1

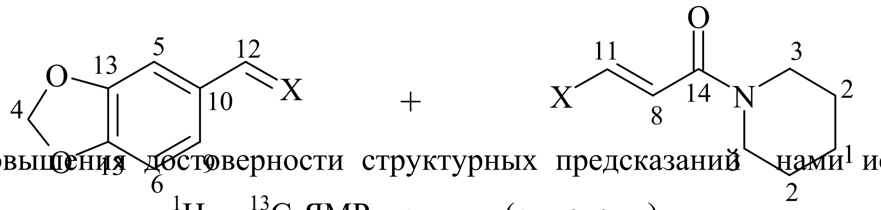
(С17)



Фрагменты, выявленные из анализа структур соединений ответов ПОИСК-13С

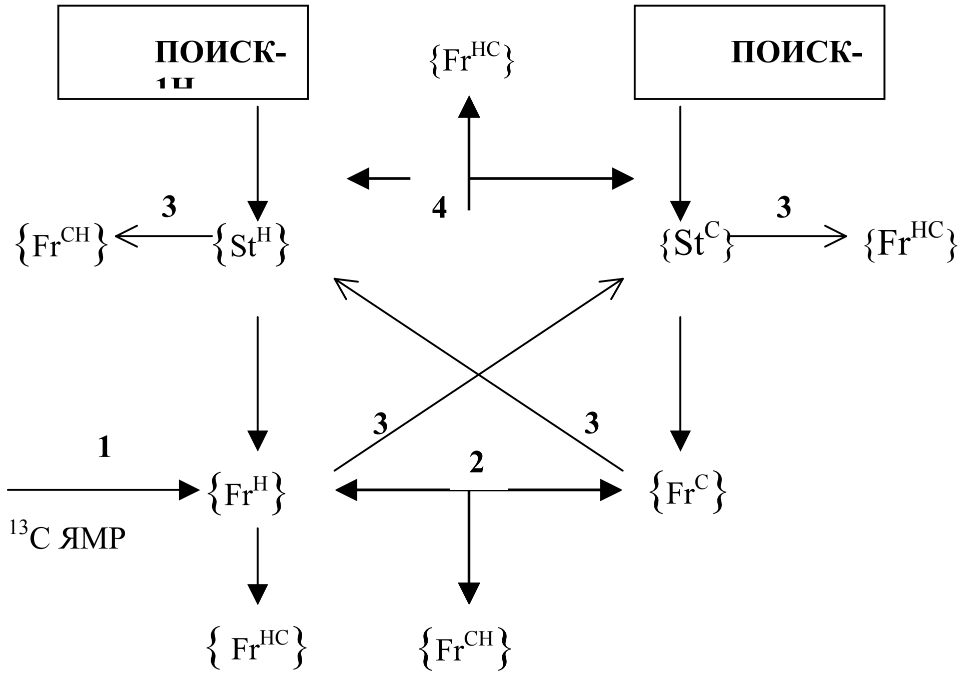


(**С16**)



(**С17**)

С целью повышения достоверности структурных предсказаний нами исследованы подходы, сочетающие анализ ^{1}H- и ^{13}C-ЯМР данных (см. **схему**).



Варианты подобных сочетаний могут быть самыми разнообразными. Простейший из

них (**1**) состоит в проверке выделенных из структур ответа ПОИСК-1Н фрагментов ($\{Fr^H\}$) на соответствие мультиплетностям сигналов спектра ^{13}С ЯМР. В качестве “фильтра” в этом случае выступает углеводородная часть фрагментарной формулы, которая надежно может быть определена из анализа спектров ^{1}Н и ^{13}С [173]. Показано, что данный прием приблизительно в два раза уменьшает число некорректных фрагментов в общем списке гипотез, выдаваемых ЭВМ. Одновременно повышается доля корректных фрагментов среди 10 первых гипотез списка с 32% (ПОИСК-1Н) до 46% (ПОИСК-1Н + спектр ^{13}С ЯМР) и растут значения параметров P_n (P_1 = 50%, P_3 = 82% и P_5 = 85%), уступая, однако, аналогичным данным для ПОИСК-13С.

Следующий достаточно простой путь (**2**) состоит в отборе одних и тех же фрагментов из списков $\{Fr^H\}$ и $\{Fr^C\}$, формируемых по результатам работы процедур ПОИСК-1Н ($\{St^H\}$) и ПОИСК-13С ($\{St^C\}$) соответственно. Обеспечивая высокую достоверность структурных предсказаний, этот подход, однако, ограничен в применении, т.к. требует наличия тождественных фрагментов в анализируемых списках. Более сложные варианты (**3**) включают подтверждение фрагментов, опознанных по одному виду спектров, структурами соединений отобранных по другому. Результатом такой проверки является список общих для структур соединений поисковых ответов фрагментов: $\{Fr^{HC}\}$ или $\{Fr^{CH}\}$. Этот вариант снимает отмеченное выше ограничение на тождественность фрагментов и расширяет круг решаемых задач. Наконец, поставленная задача может быть решена путем “перекрестного” анализа структур соединений обоих поисковых ответов с целью выделения максимально общих фрагментов (см. **4** на схеме). При этом однако, необходимо учитывать результаты отнесения отдельных узлов анализируемых структур к сигналам спектров ^{1}Н- и ^{13}С-ЯМР изучаемого соединения.

Рассмотренные выше приемы **2**, **3** и **4** имеют свои преимущества и недостатки, но общим является то, что все они приводят к списку фрагментов, удовлетворяющих двум видам спектров. Среди них подавляющее большинство (~80%) оказываются корректными, а значения параметров P_n выше, чем для фрагментов, выявляемых по отдельным видам спектров:

$$P_1 = 80\%, \quad P_3 = 93\%, \quad P_5 = 96\%. \qquad (^1\text{H} + {}^{13}\text{C ЯМР})$$

Эти результаты можно считать вполне удовлетворительными. В наших экспериментах лишь в двух случаях не получено правильного решения, т.е. среди выявленных фрагментов не оказалось ни одного, принадлежащего изучаемому соединению. Показано [193], что избежать подобных ситуаций можно только путем требования в поисковом запросе более высокой точности совпадения сравниваемых значений химических сдвигов сигналов спектров X и R (например, ± 0.3 и ± 0.07 м.д. для ^{13}С- и ^{1}Н-ЯМР соответственно). Однако

этот прием уменьшает число решаемых задач и размеры опознаваемых фрагментов. Поэтому проблему выбора между достоверностью предсказаний и размерами выявляемых фрагментов должен решать в каждой конкретной ситуации сам исследователь.

2.1.3.3. Анализ поисковых ответов с помощью корреляционных таблиц.

Результативность рассмотренных выше приемов выявления структурных фрагментов, подтвержденных двумя видами спектров, определяется наличием в соответствующих поисковых ответах однотипных структурных аналогов исследуемого соединения. Это ограничение можно устранить, если ответы, полученные по одному виду спектров (например, масс-, ИК или ПМР), анализировать с помощью таблиц спектро-структурных корреляций по другому виду спектроскопии. В этом отношении наиболее перспективно использование корреляционных таблиц ^{13}С ЯМР (далее, КТ-13С). С их помощью можно проверить все “углеродсодержащие” узлы структур соединений рассматриваемого поискового ответа на соответствие отдельным сигналам спектра ^{13}С ЯМР и выявить таким образом фрагменты, удовлетворяющие двум типам спектров исследуемого соединения. Важно отметить, что положительный результат в этом случае достигается при наличии в поисковом ответе одного близкого структурного аналога изучаемого соединения.

Подобный прием решения структурных задач апробирован с помощью КТ-13С, сформированных путем машинного анализа БД, содержащего около 5000 структур и спектров ^{13}С ЯМР органических соединений [194]. Таблицы включают список ~33000 “сферических” фрагментов (s-фрагментов), описанных в виде центрального узла (CH_3, CH_2, CH, C, CO и т.п.) и узлов его окружения, отстоящих от центрального на одну, две и три связи (например, CH_3-O- (s = 1), CH_3-O-CO- (s = 2) и CH_3-O-CO-CH_2- (s = 3)). Для каждого из фрагментов в КТ-13С указаны: среднее значение химического сдвига центрального атома, среднеквадратичное отклонение и число сигналов в спектрах БД, отнесенных к данному s-фрагменту.

На первом этапе с помощью КТ-13С для каждого соединения поискового ответа, полученного по масс- (или ИК-, или ПМР-) спектру, предсказывается модельный спектр ^{13}С ЯМР. Положения сигналов в нем определяются средними значениями химических сдвигов центральных узлов наиболее “крупных” из найденных в КТ-13С для анализируемой структуры фрагментов, а мультиплетности - типами узлов (C, CO,CN – синглет, CH, CHO – дублет, CH_2 – триплет и CH_3 – квартет). Преобразованный таким путем поисковый ответ можно рассматривать как своеобразную мини-базу данных по спектроскопии ^{13}С ЯМР, в которой вместо экспериментальных представлены модельные спектры соответствующих соединений. Поэтому последующая процедура анализа данных

такого поискового ответа сочетает в себе многие из рассмотренных в разделе 2.1.3.2. приемы. Основные отличия касаются сравнения сигналов предъявленного и модельных спектров ^{13}C ЯМР, а также оценки достоверности выявляемых фрагментов [194].

Не останавливаясь на деталях, отметим, что этапе сравнения модельный спектр рассматривается как суперпозиция распределений Гаусса, описывающих спектральное поведение - фрагментов соответствующей структуры. Это позволяет в рамках данного приближения получить для каждого сигнала экспериментального спектра вполне определенную оценку (***P***), отражающую вероятность “принадлежности” его тому или иному распределению модельного спектра, а, следовательно, и вероятность присутствия соответствующему этому распределению s–фрагмента в молекуле исследуемого соединения. Использование ряда эмпирических приемов, позволило также учесть статистические особенности корреляционных таблиц и специфику метода ЯМР. Так, например, обобщенная оценка достоверности s-фрагмента ***P*** распределяется между узлами таким образом, что центральный получает оценку всего фрагмента в целом, а дальше она распределяется между узлами окружения, включая и гетероатомные, аналогично тому, как передается электронное влияние по связям, т.е. с затуханием в большей степени на одинарных и в меньшей – на ароматических, двойных и тройных связях (см. рис.2.3).

Поскольку s–фрагменты одной структуры многократно перекрываются, то после распределения оценок по узлам каждый узел данной структуры получает итоговую оценку достоверности, состоящую из вкладов, полученных им в качестве центрального узла соответствующего s-фрагмента и узлов окружения в других фрагментах. Количественные оценки достоверности отдельных узлов позволяют выделить из структур соединений поискового ответа фрагменты, наилучшим образом удовлетворяющие предъявленному спектру. Отметим, что соблюдая требование однозначного отнесения узлов к сигналам спектра ^{13}C ЯМР, мы выделяем неперекрывающиеся фрагменты, последующее объединение которых не представляет сложной проблемы.

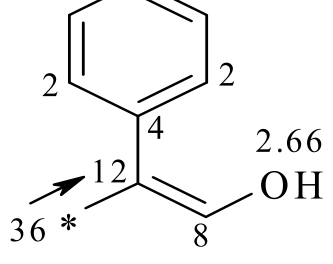
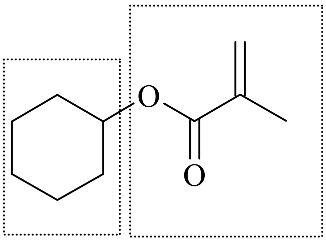


Рис. 2.3. **Пример распределения оценки достоверности s-фрагмента (*P*=36, s= 3)**

по составляющим его узлам (* - центральный узел s-фрагмента)

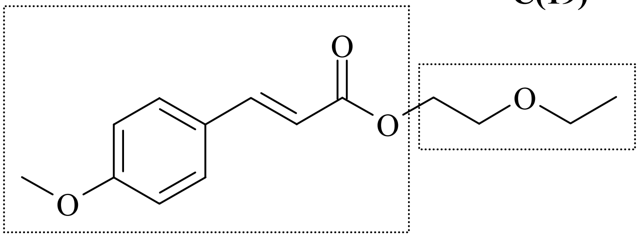
Ниже, в качестве примера, приведены структурные формулы четырех “неизвестных” соединений, в которых обведены фрагменты, выявленные таким путем из анализа поисковых ответов по масс-, ИК- и ПМР-спектрам этих соединений.

<u>ПОИСК-МС + КТ-13С</u>

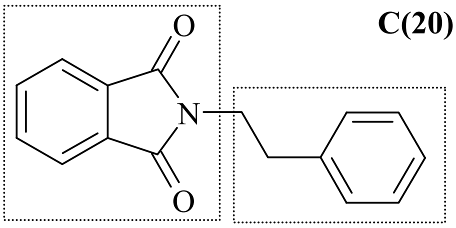


С(18)

С(19)

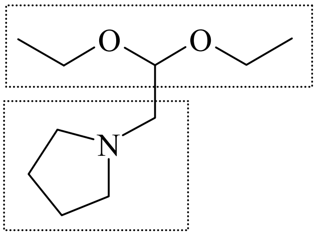


<u>ПОИСК-ИК + КТ-13С</u>



С(20)

<u>ПОИСК-ПМР +КТ-13С</u>



С(21)

Эксперименты показывают, что достоверность определяемых фрагментов практически не зависит от используемых для поиска спектров, а величины P_n не уступают приведенным в разделе 2.1.3.2 для случая совместного анализа ответов ПОИСК-1Н и -13С. Несомненным достоинством обсуждаемого подхода является также возможность моделирования спектров для проверки (построенных на основе выявленной информации) структур на соответствие экспериментальному спектру ^{13}С ЯМР.

Таким образом, разработанные методы позволяют успешно решать задачи опознания достаточно крупных структурных фрагментов изучаемого соединения, как по отдельным видам спектров, так и различным их комбинациям. В последнем случае, по-видимому, можно ограничиться двумя видами спектров (масс- и ИК; ^{1}Н- и ^{13}С-ЯМР; масс- и ^{13}С ЯМР; ИК и ^{13}С ЯМР), т.к. среди выявляемых фрагментов с высокой степенью вероятности (≥96%) оказывается, по крайней мере, один корректный. Следовательно, его использование для

построения возможных структур изучаемого соединения ведет к искомому результату.

2.1.3. Исходные данные и генерирование структурных гипотез.

Задача построения (генерирования) возможных структур соединения с помощью ЭВМ хорошо известна и получила должное решение в системах "искусственного интеллекта" [195-198]. Используемые при этом алгоритмы и программы во многом определяются исходными данными. В рамках обсуждаемого подхода удалось существенно расширить список доступной информации, который может включать: молекулярную и фрагментарную формулы, брутто-формулу формальной ненасыщенности и крупные структурные фрагменты (КСФ). Методы определения этих характеристик отличаются друг от друга. Общность состоит в том, что все они приводят к ранжированным спискам, в которых с той или иной степенью вероятности можно обнаружить данные (МФ, ФФ, БФФН и КСФ), отвечающие исследуемому соединению.

Показано, что наилучшие результаты достигаются при определении базовых характеристик соединения - молекулярной и фрагментарной формул. В этом случае искомые МФ и ФФ практически всегда генерируются на основе извлекаемой из спектров информации, а после ранжирования находятся среди первых ответов, выдаваемых исследователю. При определении крупных фрагментов возможны ситуации, когда среди выявляемых блоков не оказывается ни одного, принадлежащего изучаемой молекуле. Этого недостатка не избежать и в системах "искусственного интеллекта", оперирующих с корреляционными таблицами, в которых может быть не представлено все структурное многообразие и сравнительно небольших фрагментов. Поэтому в рамках формализованной схемы наиболее надежный путь решения структурной задачи состоит в построении исчерпывающего списка изомеров на основе сведений о МФ или ФФ. Использование последней характеристики, безусловно, предпочтительнее. Однако и в этом случае, особенно при установлении сложных соединений, может генерироваться большое число структур, последующий анализ которых крайне затруднен. Это вынуждает использовать дополнительные ограничения, в качестве которых могут выступать, например, БФФН и КСФ [199-200].

Понятно, что в плане сокращения числа генерируемых структур предпочтительнее использовать крупные фрагменты [199]. В этом случае надежный путь, приводящий к искомому результату, состоит в построении структур на основе всех не противоречащих друг другу вариантов сочетаний данных (МФ + КСФ или ФФ + КСФ), формируемых из списков соответствующих гипотез {МФ}, {ФФ} и {КСФ}. Однако это требует существенных затрат времени на многократное выполнение процедуры генерирования и проверку всех

структур на изоморфизм (тождественность). Напомним, что из анализа поисковых очень часто выделяются “пересекающиеся” фрагменты, которые и приводят к спискам, содержащим значительное число тождественных структур.

В работе [199] обсуждаются два возможных пути преодоления отмеченных недостатков. Первый состоит в отборе из первоначального списка {КСФ} лишь тех фрагментов, которые полностью “вкладываются” в более крупные, а второй - в отборе наиболее крупных фрагментов, в которые полностью “вкладываются” более мелкие. Второй вариант предпочтительнее, однако он более рискован, так как среди наиболее крупных фрагментов может не оказаться ни одного корректного. Риск в данном случае оправдывается перспективой получения сравнительно небольшого списка структур и быстрой проверкой их на соответствие экспериментальным спектрам. В этой связи, большое значение приобретает процедура генерирования, которая должна обеспечить быстрое построение всех возможных структур на основе доступных исходных данных и их сочетаний, например: МФ + КСФ, ФФ + КСФ, МФ + БФФН + КСФ. При этом необходимо предусмотреть возможность задания и априорных сведений о строении неизвестного соединения, которыми располагает исследователь.

Этим требованиям отвечает разработанная в НТЦ ХИ программа GENM [201, 202]. Во-первых, она обеспечивает генерирование только неизоморфных молекулярных графов на основе соответствующего множества помеченных вершин, которым могут соответствовать отдельные атомы и микрофрагменты типа CH_3, CO, OH и т.п.. Во-вторых, для сокращения числа генерируемых структур предусмотрена возможность задания дополнительных ограничений путем указания сведений о связывании вершин отдельных типов между собой (запрет на образование связи, максимально допустимая кратность связи между соответствующими типами верши). В-третьих, в качестве структурных ограничений могут быть заданы обязательные и запрещенные фрагменты, которым соответственно разрешается и запрещается присутствовать в структурах финального списка. При описании этих фрагментов наряду с вершинами исходного множества можно использовать вершину “универсального” типа (X - любой отличный от водорода атом), что позволяет задавать и дескрипторы формальной ненасыщенности (например, двойная связь – X=X). И, наконец, в качестве ограничений можно задавать, как непересекающиеся, так и пересекающиеся фрагменты.

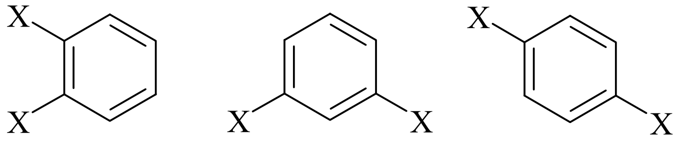
В последнее время алгоритм генерации дополнен процедурой учета желательных фрагментов, под которыми понимаются фрагменты с приписанными весами. Построенные структуры должны содержать такое подмножество желательных фрагментов, для которого сумма весов не ниже заданного порога отражения. Под множеством желательных могут

пониматься и альтернативные фрагменты. Важно отметить, что все структурные ограничения используются непосредственно в процессе построения структур, что ведет к резкому сокращению вариантов перебора исходного множества вершин и, как следствие, к быстрому получению конечных результатов. В качестве примера рассмотрим результаты генерирования структурных изомеров неизвестного соединения по следующим исходным данным [203]:

Молекулярная формула: $C_9H_{11}NO$

Брутто-формула формальной ненасыщенности: $(АК)_1(ДС)_1$

Желательные фрагменты с весами, равными единице:



Порог отражения: 1 (генерируемые структуры должны содержать хотя бы один из желательных фрагментов).

Отметим, что сведения о БФФН задают по сути дела два обязательных фрагмента: X=X и ароматическое кольцо с шестью свободными валентностями. На основе этих данных программой GENM построено 459 структур за 0.4 секунды (Пентиум-133 МГц). Для сравнения укажем, что если задать только МФ соединения, то генерируется 28 895 621 структур - за 25 минут. Если дополнительно к молекулярной формуле задано наличие ароматического кольца, то генерируется 6810 структур за 1.7 с.

Рассмотренные приемы выбора исходных данных и программные средства генерирования структур относятся к наиболее сложным ситуациям, когда информация о строении извлекается из масс- и ИК-спектров. При решении задач с использованием спектров ЯМР ситуация упрощается, поскольку выявляемые фрагменты (точнее, составляющие их узлы) имеют отнесения к отдельным сигналам анализируемого спектра (см. табл.2.4.). Это позволяет формировать наборы фрагментов, полностью согласованные с МФ, ФФ и спектром ЯМР изучаемого соединения. Это уменьшает число генерируемых структур и затраты времени. Появляется возможность контролировать процесс генерирования путем предсказания (с помощью КТ) химических сдвигов связываемых между собой узлов и сравнения этих данных с экспериментальными [204]. Показано, что этот прием позволяет существенно уменьшать число структур в финальном списке, отбрасывая заведомо неправильные структуры.

2.1.5. Ранжирование и проверка структурных гипотез на соответствие эксперименталь ным спектрам

В общем случае, на основе извлекаемой из спектров информации генерируется довольно большой список структур, из которых требуется выбрать наиболее вероятные для исследуемого соединения. Традиционно эта задача решается путем предсказания “теоретических” спектров и сравнения их с экспериментальными данными [7,46] . Такой подход получил наибольшее развитие в спектроскопии ЯМР [205-208]. Для масс-спектрометрии проблема предсказания спектров - одна из самых сложных и пока не получивших должного решения. Поэтому развиваются в основном эмпирические методы, которые по ряду причин (недостаточная обоснованность или универсальность исходных посылок, необходимость сведений о классе изучаемого соединения и т.д.) трудно считать совершенными (см.[209]). В случае ИК спектроскопии расчетные методы [210] весьма трудоемки и ограничены в применении. В связи с этим рассмотрим некоторые альтернативные пути решения поставленной задачи.

Напомним, что в результате поиска из БД отбираются соединения, подобные исследуемому по спектральным признакам. При этом структуры отобранных соединений, как правило, содержат фрагменты молекулы исследуемого соединения, а совокупность этих фрагментов нередко полностью описывает его строение. Другими словами, существует высокая вероятность того, что корреляционные связи между структурными фрагментами соединений поискового ответа и спектральными признаками предъявленного спектра носят неслучайный характер. Следовательно, достоверность генерируемых структур можно оценивать путем их сопоставления со структурами соединений поискового ответа. При этом в качестве наиболее вероятных можно рассматривать такие стуктуры, которые по совокупности признаков наиболее близки к группе соединений поискового ответа.

Рассмотрим основные особенности реализации подобного подхода на примере решения задач по масс-спектрометрическим данным [200]. Степень подобия сравниваемых пар структур в этом случае оценивается на уровне совпадения сравнительно небольших “сферических” фрагментов, представленных в виде центрального узла и узлов его окружения, отстоящих от центрального на одну связь, с указанием типов связей между смежными узлами (далее, α-фрагменты).

Например, структура соединения:

CH_3—CH_2—N—CH—CH_3
|　　|
NO　CH_3

ОПИСЫВАЕТСЯ НАБОРОМ ИЗ ШЕСТИ α-ФРАГМЕНТОВ, ПОСЛЕДНИЙ ИЗ КОТОРЫХ ВСТРЕЧАЕТСЯ ДВАЖДЫ:

$$(\overset{*}{CH_3}-CH_2-)_1(CH_3-\overset{*}{CH_2}-N<)_1(CH_2-\overset{*}{N}-\underset{NO}{\underset{|}{CH}}<)_1(\overset{*}{ON}-N<)_1(>N-\overset{*}{CH}-\underset{CH_3}{\underset{|}{CH_3}})_1(\overset{*}{CH_3}-CH<)_2$$

Звездочкой (*) помечены центральные узлы. Представленные таким образом структуры напоминают спектры, где положения линий характеризуют коды α-фрагментов, а интенсивности - число фрагментов соответствующего типа. Поэтому степень подобия двух структур (SS_{ij}) оценивается при помощи достаточно простого (использованного ранее при сравнении масс-спектров [165]) выражения:

$$SS_{ij} = 200 * W_{ij} / (W_i + W_j),$$

где W_{ij} - суммарный фактор значимости совпавших α-фрагментов в i и j структурах; W_i и W_j - факторы значимости всех α-фрагментов i и j структур соответственно. Совпавшими считаются фрагменты, в которых идентичны все составляющие их элементы; значимость α-фрагментов (W) оценивается их "размерностью", определяемой числом узлов окружения.

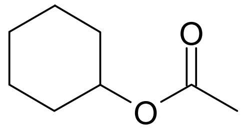
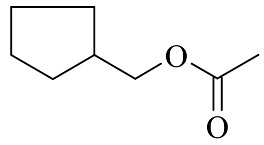
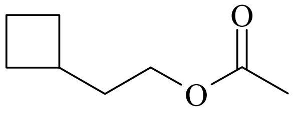
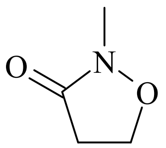
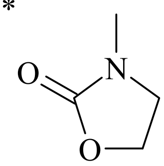
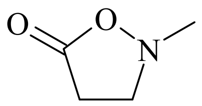
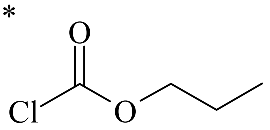
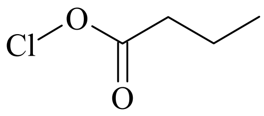
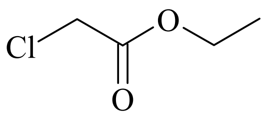
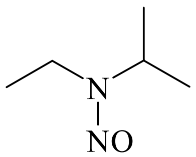
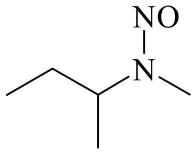
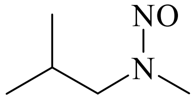
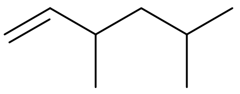
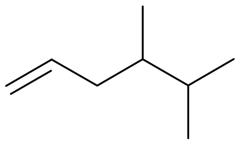
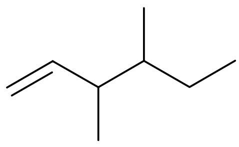
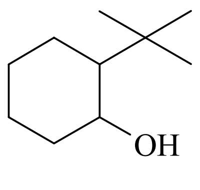
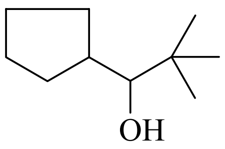
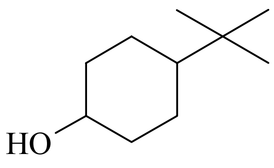
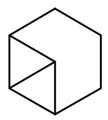
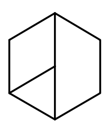
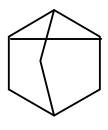
Ранжирование списка генерируемых структур $\{S^G\}$ осуществляется с помощью параметра RS_i, вычисляемого для каждой S_i^G путем ее сравнения со структурами k-первых соединений поискового ответа. По этому параметру преимущество получают структуры, которые по совокупности составляющих их α-фрагментов наиболее близки к структурам соединений поискового ответа, а спектры последних, в свою очередь, наиболее подобны спектру исследуемого соединения. Несколько примеров приведены в таблице 2.5 [200].

Искомая структура далеко не всегда находится на первом месте ранжированного списка; в наиболее неблагоприятном случае занимает 20 место. Учитывая, что эти результаты получены практически в автоматическом режиме с использованием только сведений о масс-спектрах низкого разрешения и молекулярных формул, их можно считать вполне удовлетворительными. В тех случаях, когда истинная структура не находилась на первом месте, его занимали, как правило, соединения, близкие в структурном отношении к "неизвестному".

Продолжая исследования в этом направлении, мы разработали еще один метод предварительного ранжирования структурных гипотез, в котором используются не сами структуры соединений поискового ответа, а наиболее достоверные фрагменты, выявленные из анализа этих структур [211]. Этот метод, апробированный на примере решения ряда задач по ИК-спектрам с использованием БД типа "спектр - фрагментный состав соединения", показал также хорошие результаты.

Таблица 2.5.

Результаты ранжирования генерируемых структур при определении строения шести “неизвестных” соединений по их масс-спектрам низкого разрешения

Первые структуры в ранжированном списке гипотез						N_G
1*		2		3		152
1		2*		3		20
1*		2		3		20
1*		2		3		8
1*		2*		3		11
1		2	 . . .	20*		840
1		2	 . . .	5*		10

Можно надеяться, что результативность подобных методов ранжирования будет повышаться при совместном использовании поисковых ответов, полученных по масс-и ИК-спектрам. Однако результаты ранжирования в любом случае будут носить предварительный характер, и будут, по сути дела, определять порядок последующей более тщательной проверки (но уже ограниченного числа) структур на соответствие экспериментальным

спектрам. Эта проблема может быть также решена с помощью банков данных типа “структура-спектр” [14, 98].

Предлагаемый подход весьма прост и состоит в выполнении следующих операций:

- отбор из БД соединений, структуры которых содержат, по возможности, крупные и специфические в спектральном отношении фрагменты проверяемой молекулы;
- статистический анализ спектров отобранных групп соединений с целью выявления характеристических для соответствующих фрагментов спектральных признаков;
- построение на основе выявленных признаков модельного спектра и сравнение его со спектром исследуемого соединения.

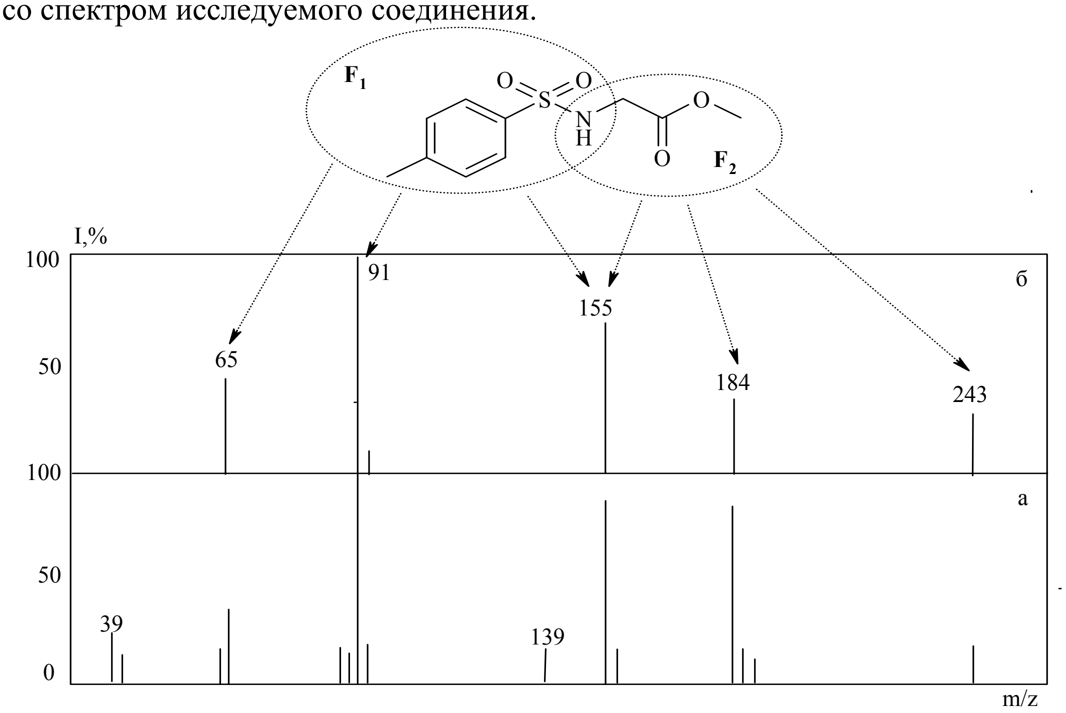
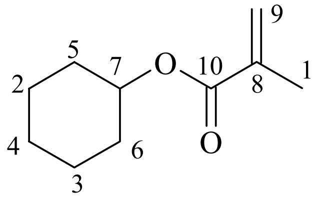


Рис.2.4. **Экспериментальный (а) и моделированный (б) масс-спектры метилового эфира N-(п-тозил)глицина**

В качестве примера на рис.2.4 приведены экспериментальный (а) и моделированный (б) масс-спектры одного из соединений, взятые из работы [14]. Модельный спектр построен на основе анализа спектров соединений базы данных, содержащих фрагменты $\mathbf{F_1}$ (область глубокого распада молекулы) и $\mathbf{F_2}$ (область молекулярного иона и ближайших к нему осколочных). Успех моделирования во многом определяется опытом исследователя при выборе “ключевых” фрагментов и наличием в БД достаточного числа соединений, содержащих эти фрагменты. Строгость выполнения аддитивных схем может обеспечить более высокую эффективность этого приема в ЯМР- и ИК-спектроскопии.

Для моделирования спектров ^{13}C ЯМР, предпочтительнее, по-видимому, использовать не БД, а сформированные на их основе корреляционные таблицы, так как в этом случае возможно решение задачи без участия исследователя [194]. Наличие в таблицах сведений о спектральном поведении s-фрагментов, описанных от первой до третьей (или более дальних) сфер окружения центральных атомов углерода, позволяет моделировать спектры практически любых проверяемых структур. При этом использование более крупных s-фрагментов приводит к более “точным” модельным спектрам. Проиллюстрируем сказанное лишь одним примером предсказания спектра ^{13}C ЯМР с помощью сформированных машинным путем КТ, содержащих s-фрагменты до четвертой сферы окружения центральных атомов (s = 1, 2, 3, 4):

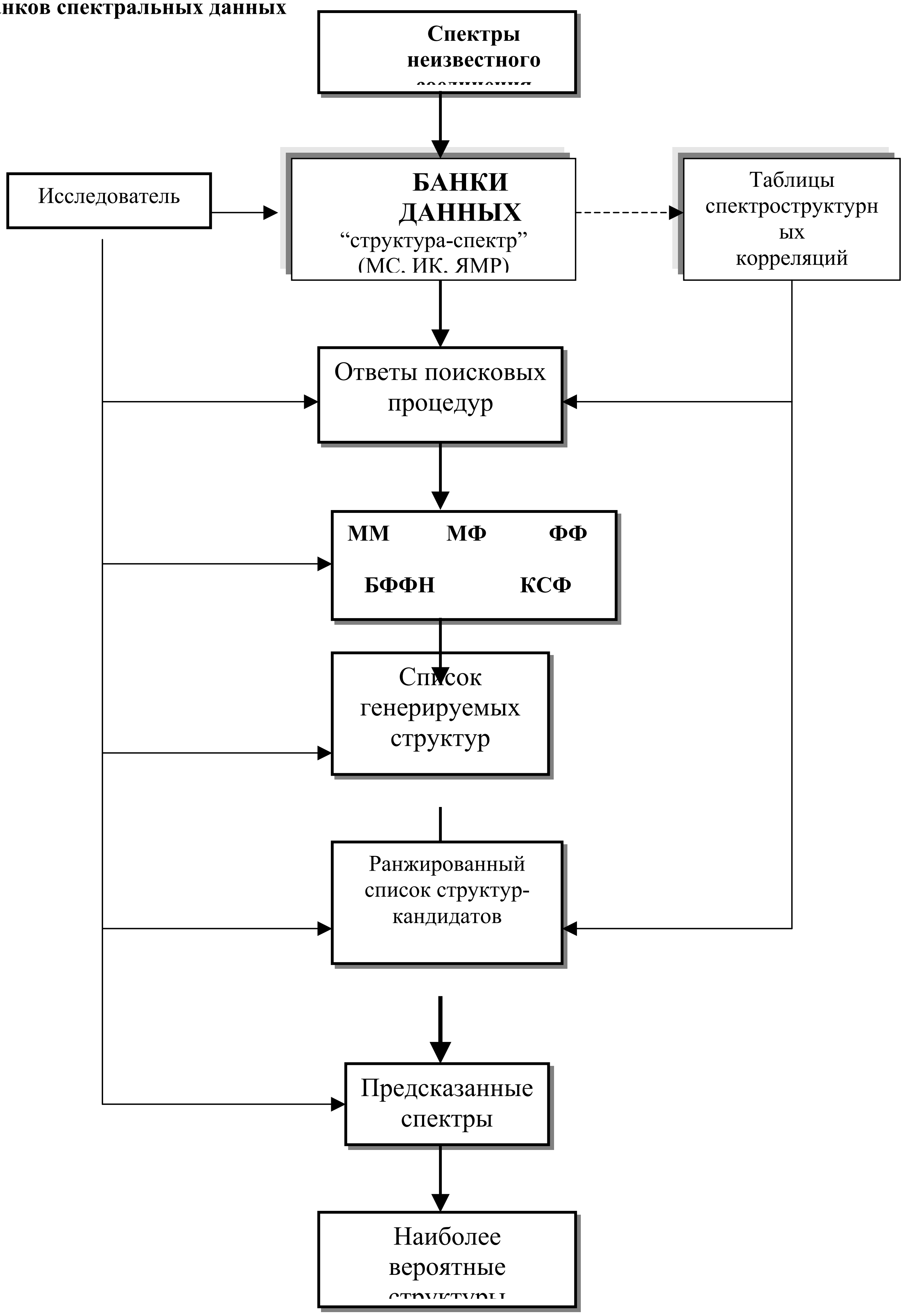


(C18)

Номер резон. атомов углерода	Предсказанные значения химических сдвигов с использованием различных s-фрагментов				Наблюдаемые в спектре сигналы
	s = 1	s = 2	s = 3	s = 4	
1	21.2	18.6	18.2	18.3	18.3/к
2	29.3	25.3	23.8	23.8	23.8/т
3	29.3	25.3	23.8	23.8	23.8/т
4	29.3	29.5	23.8	23.8	25.7/т
5	33.6	31.9	31.7	31.6	31.7/т
6	33.6	31.9	31.7	31.6	31.7/т
7	76.0	72.7	73.1	72.6	72.5/т
8	137.5	136.7	137.3	137.3	137.2/с
9	117.8	123.7	125.0	124.5	124.6/т
10	165.0	166.5	166.5	166.5	166.6/с

В последнее время нами разработан новый метод предсказания спектров ^{13}C ЯМР, который, с одной стороны, использует все преимущества описанного выше (простота и быстродействие), а с другой - позволяет решать задачи при отсутствии в КТ данных о химических сдвигах атомов углерода с тем или иным вариантом окружения, представленном в заданной структуре [204].

Рис.2.5. **Общая схема решения структурно-аналитических задач с помощью банков спектральных данных**



В этом случае делается попытка использовать данные о родственных видах окружения, полученных путем по парных сравнений всех видов окружения атомов углерода, представленных в КТ-13С. Для каждого найденного варианта окружения вычисляются ожидаемое смещение химического сдвига и “мера недоверия”, которые затем используются при подсчете ожидаемого значения химического сдвига атома углерода в искомом варианте окружения. Показано, что этот метод (названный нами адаптивным) позволяет в общем случае получать более “точные” спектры, что находит свое отражение на результатах ранжирования структурных гипотез по мере совпадения предсказанных для них спектров и экспериментальных данных ^{13}C ЯМР. Необходимо отметить и несколько иные, основанные на использование баз спектро-структурных данных, приемы моделирования спектров [212-214].

Таким образом, совокупность разработанных методов позволяет решать весь комплекс проблем, возникающих при установлении строения неизвестного соединения, от определения его молекулярной массы до списка наиболее вероятных структурных формул (см. рис.2.5). В зависимости от исходного набора спектров возможны различные варианты решения задач с использованием банков данных и таблиц спектроструктурных корреляций. Заметим, что рассмотренные в этом разделе методы и приемы не отвергают существующие, а расширяют и дополняют их, особенно, в части определения базовых характеристик соединения (МФ и БФФН), выявления крупных структурных фрагментов (КСФ), предварительного ранжирования структурных гипотез и моделирования спектров. Это открывает реальную возможность создания систем нового поколения, сочетающих достоинства двух основных подходов к установлению строения неизвестных соединений, основанных на использовании банков спектро-структурных данных и баз знаний по различным видам молекулярной спектроскопии.

2.2. Информационно-аналитические системы по молекулярной спектроскопии

При решении структурно-аналитических задач по спектральным данным (МС, ИК, ЯМР) традиционными методами исследователь выполняет большой объем работы. Это анализ спектров, просмотр справочной литературы, формулирование гипотез о строении соединения, проверка их на соответствие экспериментальным спектрам и т.п. К настоящему времени созданы компьютерные средства, облегчающие выполнение многих из отмеченных операций (см., например [215-221]). Большинство разработок, однако, существует в виде отдельных программ и систем, достаточно полный набор которых доступен лишь

ограниченному кругу исследователей. Это обстоятельство требует создания многофункциональных систем, в рамках которых можно было бы получить ответы на большинство или, по крайней мере, основную часть вопросов, возникающих в процессе установления строения соединения по спектральным данным.

При разработке таких систем следует иметь в виду, что в каждом конкретном случае нельзя предугадать логику исследователя при выборе наиболее оптимальных путей решения той или иной задачи. Нельзя предусмотреть последовательность, вид данных, необходимых на отдельных этапах этого пути и порядок использования программ для наиболее быстрого достижения результата. Поэтому при разработке подобных систем необходимо учитывать не только многофункциональность, но и требуется модульная организация, позволяющая с одной стороны разработчику легко расширять систему, а с другой, пользователю самому определять путь решения тех или иных задач с использованием программных средств, реализованных в системе.

Наш опыт показывает, что максимальный эффект применения компьютерной системы достигается в том случае, если ей поручается обработка большого объема информации по формализованным правилам, например, поиск в базе данных спектральных аналогов, генерирование изомеров на основе молекулярной формулы, статистический анализ выборки спектров и т.п. Во всех случаях, однако, после выполнения соответствующих операций окончательное решение должен принимать исследователь, т.е. компьютер лишь помогает на наиболее трудоемких этапах перебрать и оценить все варианты возможных решений. Поэтому разрабатываемые системы должны представляет собой не совокупность полностью автоматизированных процедур, а набор программных средств, умелое использование которых облегчает и ускоряет решение задач. В соответствии с общей схемой (см. рис.2.5), процедуры подобных систем можно разделить на две большие группы – поисковые и аналитические. Первые предназначены для отбора из БД информации, релевантной запросу, а вторые – для решения задач на основе анализа этой информации. Поисковые процедуры должны, в свою очередь, обеспечить работу со спектральными и структурными разделами БД, а аналитические – решать задачи по отдельным видам спектров и их комбинациям с использованием результатов поиска и корреляционных таблиц. При этом исследователь должен иметь возможность принимать участие на всех этапах решения задачи. Это касается ввода исходных данных, просмотра и корректировки результатов работы поисковых процедур, выбора данных для аналитических процедур и т.п.

В настоящее время в рамках общей схемы и рассмотренных выше принципов построения информационно-аналитических систем разработаны две системы, предназначенные для оказания исследователю помощи при решении задач только по масс-

спектрам (КОМПАС-МС) [98, 222] и набору спектров (ХимАрт) [223,224]. Рассмотрим основные особенности этих систем.

2.2.1. КОМПАС-МС – информационно-аналитическая система по масс-спектрометрии

Информационную основу системы составляет банк данных, содержащий сведения о масс-спектрах низкого разрешения и структурных формулах порядка 50000 органических соединений. В нем представлены преимущественно полные спектры из отечественных и зарубежных источников, записанные на приборах с разверткой по магнитному полю и энергии ионизирующих электронов 50-70 эВ. Решение предусмотренных в системе задач осуществляется при помощи процедур, каждая из которых включает набор операций, выполняемых программными модулями с участием исследователя. Результаты работы отдельных процедур записываются в архивный файл и доступны для любых других процедур системы. В рабочем варианте системы реализовано пять поисковых (ПОИСК-1, -2, -А, -В, -АВ) и пять аналитических процедур (ММБФ, ФРАГ, ГЕНС, СТАТ-А и СТАТ-В).

ПОИСК-1 и **–2** предназначены для отбора из базы данных соединений, удовлетворяющих заданному набору спектральных (m/z ионов, массы первичных нейтральных потерь и интенсивности соответствующих пиков) и "химических" (молекулярная масса, элементный состав и структурный фрагмент) признаков соответственно. С помощью этих процедур можно решать широкий круг задач справочного характера, включая идентификацию соединения по ограниченному набору признаков. Реализованные в рамках ПОИСК-2 возможности позволяют отбирать из БД соединения одного гомологического ряда; соединения, содержащие определенный структурный фрагмент или функциональную группу; соединения с заданным типом "геометрии" (например, циклы с указанным числом вершин). Понятно, что анализ отбираемых при этом спектров способен подсказать закономерности спектрального поведения группы соединений и может оказаться полезным при проверке структур на соответствие экспериментальному спектру (см. ниже).

ПОИСК-А, -В и **-АВ** предназначены для отбора из БД соединений, спектры которых наиболее близки спектру Х по пикам ионов, пикам "первичных нейтральных потерь" и совокупности этих признаков. Широкий ассортимент способов сопоставления спектров и вычисления факторов спектрального подобия обеспечивает успешное решение задач идентификации и отбора структурных аналогов неизвестного соединения в самых разнообразных ситуациях, связанных с чистотой образца и "качеством" анализируемого спектра.

ММБФ предназначена для определения молекулярной массы и молекулярной формулы соединения по масс-спектру низкого разрешения. В основе процедуры лежат оригинальные алгоритмы [148], позволяющие в подавляющем большинстве случаев получать правильные сведения о МФ соединения, в том числе и в ситуациях, когда в предъявленном спектре отсутствуют пики молекулярных ионов.

ФРАГ предназначена для выявления крупных структурных фрагментов неизвестного соединения. В рамках этой процедуры реализованы два метода решения данной задачи [192]. Первый метод основан на “перекрестном” анализе структур соединений поискового ответа с целью нахождения максимально общих фрагментов, а второй – на декомпозиции структур для выявления фрагментов, удовлетворяющих первичной модели фрагментации органических молекул по действием ионизирующего напряжения. Достоверность выявленных фрагментов оценивается путем сравнения спектра неизвестного соединения со “спектральными откликами” каждого фрагмента, формируемыми из анализа спектров соединений поискового ответа, содержащих данный фрагмент. Наличие принципиально различных методов анализа позволяет выбирать в каждой конкретной ситуации наиболее подходящий с точки зрения исследователя путь решения поставленной задачи.

ГЕНС и **РАНС** предназначены для генерирования на основе выявляемой информации возможных структур изучаемого соединения и их ранжирования с целью выбора наиболее вероятных (см. разделы 2.1.4 и 2.1.5). Результаты представляются в виде рисунков структурных формул, что существенно облегчает визуальный анализ полученных данных.

СТАТ-А и **-В** предназначены для анализа спектров с целью выявления характеристических (наиболее часто встречающихся) признаков: пиков ионов и пиков первичных потерь, связанных с абсолютными и относительными положениями линий в спектре соответственно. Эти процедуры в сочетании с ПОИСК-2 полезны при проверке возможных направлений первичной фрагментации молекулярных ионов, выявлении спектро-структурных корреляционных связей в ряду отдельных классов соединений, моделировании масс-спектров, т. е. предсказание положений и интенсивностей пиков, но не процессов распада (см. рис.2.4).

Можно видеть, что сравнительно небольшой набор процедур предоставляет исследователю широкие возможности при поиске ответов на разнообразные вопросы, возникающие при анализе масс-спектрометрических данных. Более того, легко заметить, что в рамках системы КОМПАС-МС предпринята попытка создания “замкнутой технологической” схемы решения структурных задач методами масс-спектрометрии низкого разрешения: от спектра к структурной гипотезе и наоборот. Система реализована на IBM PC в среде MS DOS и доступна широкому кругу пользователей. Она может быть достаточно легко включена в состав программного

обеспечения спектральных комплексов для реализации следующей автоматизированной схемы анализа:

ВЕЩЕСТВО → МАСС-СПЕКТРОМЕТР → КОМПАС-МС → ГИПОТЕЗЫ О СТРОЕНИИ

Заметим, что система КОМПАС-МС не только не уступает известным коммерческим системам в области масс-спектрометрии (MSSS, STIRS, PBM, SISCOM)., но по ряду показателей превосходит их.

2.2.2. ХимАрт – мультиспектральная система решения структурных задач

Система включает базы данных типа "структура-спектр" по масс-спектрометрии (5000 записей), ИК спектроскопии (35000), спектроскопии ^{1}H ЯМР (44000) и ^{13}C ЯМР (27000). Структурные формулы соединений в этих базах данных представлены в виде "глубоких древесных" кодов соответствующих молекулярных графов [225]. Такой способ представления обеспечивает малые затраты памяти при хранении структур и быстрый последовательный подструктурный поиск. Например, файл из 50000 структур со средним числом вершин, близким к 17, занимает на диске около 1.1 Мбайт.

При работе с системой используются сформированные машинным путем для каждой базы данных классификаторы структур, обеспечивающие прямой доступ к структурно-родственным соединениям [226], и таблицы спектро-структурных корреляций ^{13}C ЯМР. Классификатор представляет собой лексикографически упорядоченный список строго канонических широких "древесных" кодов структур относительно каждой вершины во всех структурах коллекции.

Программное обеспечение системы реализовано в виде набора независимых Windows приложений, позволяющих выполнять следующие основные операции:

- поиски в базах данных соединений, спектры которых наиболее близки к предъявленным;
- анализ результатов поиска по отдельным видам спектров (масс-, ИК, ^{1}H ЯМР и ^{13}C ЯМР) с целью выявления структурных фрагментов изучаемого соединения;
- совместный анализ результатов поиска по различным видам спектров (масс- + ИК, масс- + ^{13}C ЯМР, ИК + ^{13}C ЯМР) с целью выявления фрагментов, удовлетворяющих двум видам спектров;
- анализ результатов спектральных поисков с помощью корреляционных таблиц ^{13}C ЯМР;
- анализ масс-спектра низкого разрешения и спектров ЯМР (^{1}H, ^{13}C) с целью

определения молекулярной массы и брутто-формулы соединения;

- генерирование на основе молекулярной формулы и ограничений (обязательные, запрещенные и желательны фрагменты) исчерпывающего списка структурных изомеров;
- проверка генерируемых структур на соответствие спектрам ^{1}H и ^{13}C ЯМР;
- предсказание с помощью корреляционных таблиц спектров ^{13}C ЯМР;
- поиск в базе данных соединений по структурной формуле и структурным фрагментам;
- поиск структурных аналогов изучаемого соединения относительно отдельных вершин;
- поиск наиболее близких структурных аналогов;
- анализ результатов структурных поисков с целью построения модельных спектров соединений заданного строения.

С помощью этих приложений можно решать широкий круг разнообразных задач химической практики: идентификация ранее описанных соединений, определение молекулярной формулы и структурных фрагментов неизвестного соединения, проверка структурных гипотез на соответствие экспериментальным спектрам. В настоящее время в рамках системы ХимАрт реализованы три специальных приложения **MS**, **MS+13C** и **MS+IR+13C**, ориентированных на установление строения неизвестного соединения по его масс-спектру низкого разрешения [192], масс- и ^{13}C ЯМР-спектрам [203-204], масс-, ИК- и ^{13}C ЯМР-спектрам [224] соответственно. Каждое из этих приложений включает в себе как единое целое многие из перечисленных выше программных средств, а результатом работы является ранжированный список наиболее вероятных структур неизвестного соединения. Понятно, что результативность решения задачи возрастает по мере увеличения числа используемых спектров. Рассмотрим в качестве примера приложение **MS+13C**, которое может последовательно выполнть следующие операции:

- поиск в БД-МС соединений, спектры которых наиболее близки масс-спектру неизвестного соединения;
- поиск в БД-13С ЯМР соединений, спектры которых наиболее близки спектру ^{13}C ЯМР неизвестного соединения;
- анализ ответа ПОИСК-МС и спектров ^{13}C ЯМР с целью определения молекулярной формулы неизвестного соединения;
- предсказание с помощью КТ-13С для соединений ответа ПОИСК-МС спектров ^{13}C ЯМР;
- сравнение предсказанных спектров с экспериментальными данными с целью выделения из структур соединений ответа ПОИСК-МС фрагментов, удовлетворяющих МФ и спектру ^{13}C ЯМР неизвестного соединения;

- выделение фрагментов из структур соединений ответа ПОИСК-13С;
- составление наборов фрагментов, полностью согласованных с МФ и ^{13}С ЯМР спектром неизвестного соединения, и генерирование на их основе структурных гипотез;
- предсказание для генерированных структур спектров ^{13}С ЯМР и сравнение их с экспериментальными данными;
- ранжирование структурных гипотез по мере похожести сравниваемых спектров.

В качестве примера на рис.2.6 приведены структурные формулы 10 “неизвестных” соединений, которые были определены с помощью **MS+13C** [203]. Можно видеть, что в большинстве случаев истинная структура занимает первое место в ранжированном списке гипотез. Обращает на себя внимание и сравнительно небольшое число генерируемых структур.

Эти результаты в целом положительно характеризуют реализованный метод решения структурных задач, и он, с нашей точки зрения, может быть рекомендован для практического использования. Вместе с этим следует отметить, что ряд задач был решен не полностью в автоматическом режиме, а при активном участии исследователя. Это касается в первую очередь выбора значений параметров, управляющих работой поисковых и аналитических процедур. Наши исследования показали, что подобрать оптимальные значения параметров, обеспечивающие получение удовлетворительных результатов при решении самых разнообразных структурных задач, практически невозможно. Поэтому пользователю предоставляется возможность изменять значения практически всех (выставленных в системе “по умолчанию”) параметров, чтобы найти приемлемый вариант. В связи с этим хочется еще раз подчеркнуть, что разрабатываемые нами компьютерные средства предназначены в основном для оказания исследователю помощи на наиболее трудоемких этапах решения структурной задачи, в то время как выбор пути и окончательного ответа остается за ним. Наш опыт и опыт других исследователей [46] показывает, что именно разумное сочетание интеллекту альных возможностей исследователя и вычислительных ресурсов компьютера позволяет создать системы, способные решать задачи, недоступные исследователю и компьютеру порознь

В заключение данного раздела отметим, что рассмотренные выше методы решают задачи, исходя из наиболее простых и доступных видов спектров. Понятно, что привлечение современных спектральных данных (особенно двухмерных спектров ЯМР) будет способствовать повышению результативности методов и сложности решаемых задач.

.

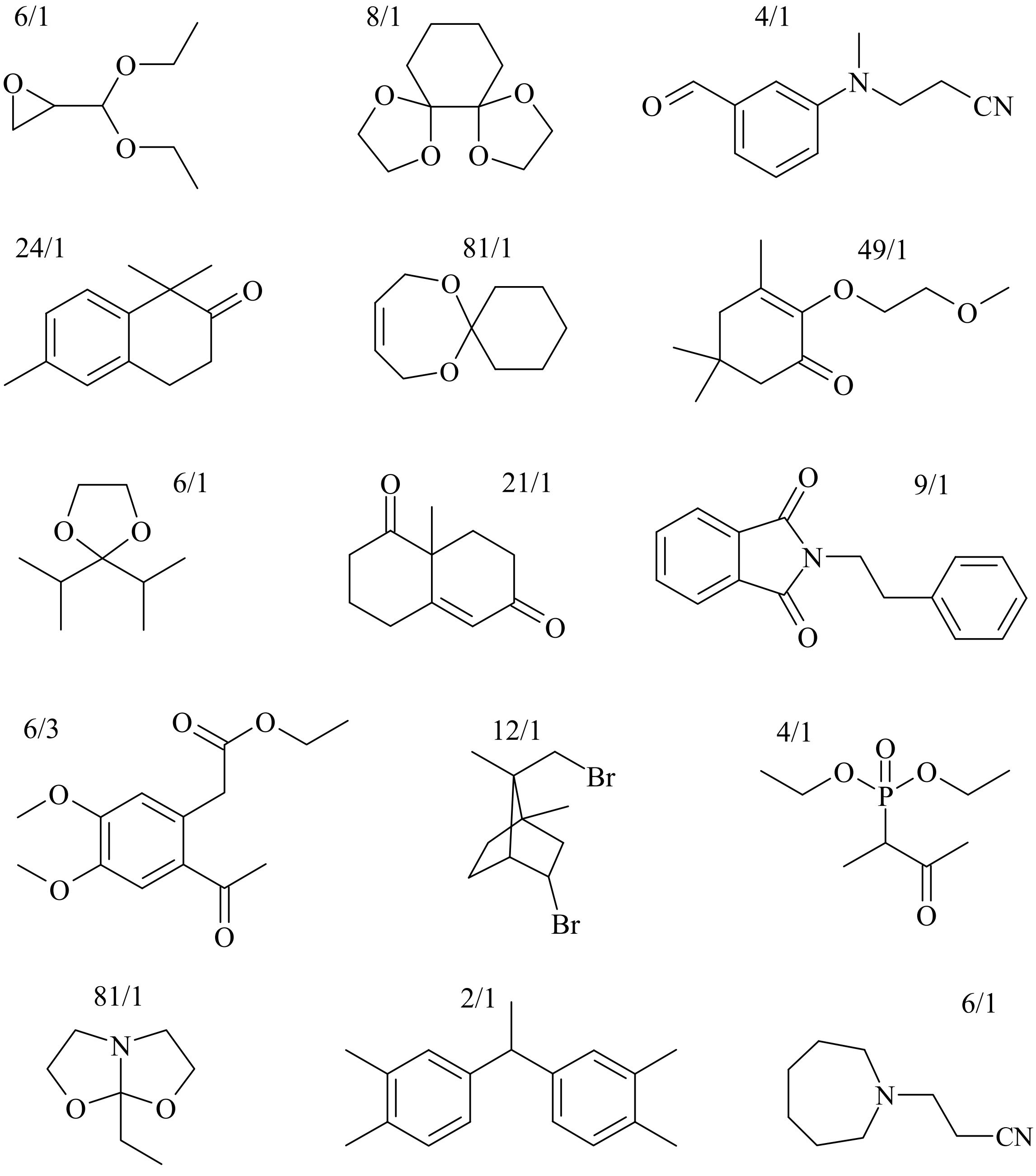


Рис. 2.6. **Примеры решения структурных задач с помощью приложения MS+13C системы ХимАрт.** В числителе – общее число генерируемых структур, в знаменателе – место определяемой структуры в ранжированном списке гипотез.

Глава 3

КОМПЬЮТЕРНЫЙ КАЧЕСТВЕННЫЙ АНАЛИЗ СМЕСЕЙ ПО ХАРАКТЕРИСТИКАМ УДЕРЖИВАНИЯ ПРИ ХРОМАТОГРАФИЧЕСКОМ РАЗДЕЛЕНИИ. ПРОБЛЕМА ИДЕНТИФИКАЦИИ КОМПОНЕНТОВ С ЗАДАННОЙ НАДЕЖНОСТЬЮ

3.1. Качественный хроматографический анализ по характеристикам удерживания

При исследовании качественного состава смесей естественно вначале попытаться выделить компоненты в чистом виде, а затем опознавать их по спектрам того или иного типа так, как описано в первых главах этой книги. Вместе с тем, разделение смесей органических веществ обычно проводят хроматографическим методом и само поведение компонентов при хроматографировании позволяет опознавать их, даже не используя спектральные данные. Далее рассматривается именно такой способ идентификации. Аналитические возможности компьютерного спектрального анализа неразделенных смесей будут рассмотрены в следующей главе.

Компьютерный качественный анализ смесей (с предварительным хроматографическим разделением или без него) требует строгих количественных оценок, а следовательно, нуждается в специальной терминологии. Введем некоторые термины.

Компонент *(component)* - определенный вид молекул (атомов, фаз и т.п.), наличие которого проверяется в пробе в ходе анализа. Лишь часть из М проверяемых компонентов реально присутствует в исследуемой пробе. Будем обозначать компонент символом X_i. Компонентом в принципе можно считать и любую совокупность более простых компонентов, имеющих общие признаки (например все фенолы, имеющие идентичные спектры поглощения, или сумму изомерных углеводородов C_6H_{14}, имеющих одно и то же время удерживания в хроматографической колонке).

Испытание *(test)* - измерение или визуальная оценка некоторого свойства пробы в строго определенных, заранее заданных условиях; например, измерение интенсивности излучения на длине волны, характерной для излучения X_i. Другой пример - измерение теплопроводности газа на выходе из хроматографической колонки (сигнал катарометра) в определенный момент времени после ввода пробы. Наиболее устойчивые признаки получают при бинарном кодировании измеряемой величины, когда результату испытания

можно приписать лишь два значения: 1 и 0. Например, в спектре пробы (или на хроматограмме пробы) соответствующий пик либо обнаруживается, либо нет. Результат испытания может быть признаком присутствия (или отсутствия) X_i в пробе.

Признаком присутствия компонента *(attribute, feature)* при бинарном кодировании является превышение критического уровня физической величины, измеряемой при единичном испытании. При других способах кодирования признаком может быть совпадение пробы и эталона по $I_{отн}$ - относительной интенсивности измеряемой величины. Такие признаки широко используют при опознании чистых веществ (в том числе поочередно выходящих из хроматографической колонки), но эти признаки менее устойчивы к влиянию посторонних веществ, а потому менее эффективны в анализе неразделенных смесей. Так, относительную интенсивность спектральных линий при опознании компонентов используют довольно часто, но лишь в качестве дополнительного поискового признака. Основным же признаком является сам факт наличия некоторой линии (пика) в спектре пробы - на данной длине волны, или при данном значении *m/z*, или при данном значении химического сдвига, и т.п.

Каждый компонент характеризуется своим набором признаков. Для обнаружения X_i используют l_i его признаков, например, l_i спектральных линий с разными длинами волн. Совокупность однотипных признаков X_i , включенных в БД, будем считать его *эталонным спектром,* независимо от природы признаков. В таком же смысле можно понимать и термин "*спектр пробы*".

Признак характеризуется величинами, не зависящими от концентрации (положение спектральной линии, интенсивность пика по отношению к другому пику того же X_i , индекс хроматографического удерживания, потенциал полуволны или иная константа D_{ij}). Другая характеристика признака - минимальная концентрация (C_{ij}) компонента в пробе, обеспечивающая положительный результат испытания (по j-ому признаку X_i). Часто признак с точностью до погрешности измерения D_{ij} является общим для нескольких (R) предполагаемых компонентов пробы (*межэталонные наложения*).

Разработчики ИПС для анализа смесей обычно принимают следующие допущения, относящиеся к признакам:

1. Все М предполагаемых компонентов пробы и все их аналитические признаки известны, компоненты совместимы и могут изучаться порознь.

2. Все признаки независимы, то есть наличие других компонентов смеси не мешает проявлению признаков X_i и не влияет на константы D_{ij}. Последнее возможно, если компоненты не реагируют друг с другом, или их взаимодействие не сказывается на результатах испытаний.

Указанные ограничения нередко не выполняются (неполные базы данных, неустойчивые признаки, неаддитивные смеси), но на данном этапе развития теории идентификации эта модель представляется неизбежной.

Проверка (*matching, comparison*) - серия испытаний пробы по заранее отобранным признакам i-го компонента. В ходе проверки используются таблицы качественных реакций, атласы эталонных спектров, наборы индексов Ковача и другие банки данных по предполагаемым компонентам, в т.ч. компьютерные БД. Предполагается, что по любому признаку результат единичного испытания может иметь лишь два исхода: совпадение свойств эталона и пробы, либо их несовпадение, а испытания с неясным исходом исключены. Проверка пробы на признаки X_i дает n_i совпадений и (l_i - n_i) несовпадений, что позволяет с определенной вероятностью считать этот компонент идентифицированным или, наоборот, отсутствующим в пробе.

В ходе хроматографического анализа проверяют совпадение характеристики удерживания пика на хроматограмме пробы с характеристиками удерживания эталонных веществ (предполагаемых компонентов пробы) в тех же условиях разделения [11][1]. Тот же прием используют при компьютерной расшифровке хроматограмм. Если t – некоторая характеристика проверяемого пика (время удерживания, удерживаемый объем, индекс Ковача и т.п.), t_X – значение той же характеристики для эталонного вещества X, а d – заданный пользователем критерий, то вывод о предположительном присутствии X делают при $|t- t_X| < d$, в противном же случае говорят о необнаружении X на данном концентрационном уровне.

Хроматографическая идентификация возможна при полном разделении компонентов смеси и отсутствии химических превращений при разделении. Но даже при выполнении этих очевидных требований использование для каждого X_i лишь одного поискового признака делает идентификацию по характеристикам удерживания (при использовании одной колонки) в принципе менее достоверной, чем опознание веществ по их спектрам, когда используют десятки, а то и сотни признаков, например, длины волн и относительные интенсивности ряда линий опознаваемого элемента.

Ошибки идентификации по характеристикам удерживания могут быть связаны с несколькими факторами. Три из них представляются наиболее важными:

- Случайные сдвиги пиков. Они характерны для всех вариантов хроматографического анализа (ГЖХ, ВЭЖХ и т.п.). Идеального совпадения значений t не бывает даже в повторных опытах. Сдвиг пика пробы, выводящий его за пределы “окна” $t_X \pm d$, может объясняться неконтролируемыми колебаниями скорости газа-носителя или иной

[1] К сожалению, иногда процесс проверки полностью сводят к этой процедуре, которая всегда необходима, но во многих случаях недостаточна для достоверной идентификации.

подвижной фазы (ПФ), нестабильностью температуры, невоспроизводимостью ввода пробы, погрешностями регистрации хроматограммы и другими причинами [227].

- Межлабораторная невоспроизводимость и непостоянство значений t_X. Табличные значения t_X не являются физическими константами, они несколько меняются при переходе от одной колонки к другой, с такой же неподвижной фазой (НФ). Характеристики удерживания постоянны только в определенном интервале варьируемых параметров и только для хорошо разрешенных пиков идеальной (гауссовской) формы. Ценность этих поисковых признаков снижается из-за зависимости удерживания от объема и даже от состава разделяемой пробы [228]. Известны случаи, когда изменение в 3 раза объема пробы сдвигало индекс удерживания X на полярной НФ на 30-35 единиц, хотя прибор, колонка и методика анализа оставались неизменными. Даже при постоянном объеме пробы изменение в 2-3 раза отношения концентраций X и ближайшего к нему н-алкана иногда приводило к сдвигу индекса удерживания X на 5-10 единиц [229]. Конечно, эти случаи – скорее редкое исключение, чем правило, но следует учесть, что разные компоненты сложной пробы иногда различаются по индексу всего на 1-2 единицы [230].
- Случайные совпадения характеристик удерживания разных X, в том числе не имеющих общих структурных фрагментов и относящихся к разным классам. Значимость фактора межэталонных наложений нарастает при усложнении состава пробы.

Для увеличения достоверности хроматографической идентификации проводят повторную проверку совпадения характеристик пробы и эталона в различных условиях, например, при разных температурах, разном составе элюента или с использованием нескольких НФ. Совокупность характеристик удерживания одного и того же X на неподвижных фазах разной полярности (или, шире, в разных условиях разделения) принято называть *хроматографическим спектром X.* “Спектральный” подход к хроматографической идентификации, предложенный Франсом, развит школой М.С.Вигдергауза. Еще более мощное средство - применение селективных (“идентифицирующих”) детекторов и дополнительных поисковых признаков [231]. Наиболее надежные результаты дает сочетание опознания по характеристикам удерживания с масс-спектральной (примеры см. в главе 1), ИК-спектрометрической или химической идентификацией компонентов после их разделения. Именно поэтому в практике анализа все шире применяют хроматографы с масс-спектрометрическим детектированием.

Давно идет эмоциональная и не слишком плодотворная дискуссия: одни хроматографисты считают идентификацию органических веществ по характеристикам

удерживания в принципе ненадежной, а иногда даже опасной [232], а другие, не менее известные и опытные, создают банки данных (БД) по характеристикам удерживания [233,234], разрабатывают методики качественного анализа реальных объектов с применением БД [235] и компьютерные программы для опознания компонентов по мере их выхода из колонки [236]. Тем не менее надежность хроматографической идентификации до последнего времени специально не изучалась (несмотря на неоднократные призывы [5, 270]); не существует даже общепринятых способов ее объективной количественной оценки. Поэтому при рассмотрении в настоящей книге аналитических возможностей метода компьютерной хроматографической идентификации основное внимание будет уделено именно проблеме надежности. Все примеры относятся к одному из вариантов - методу газожидкостной хроматографии (ГЖХ), но соответствующие закономерности и выводы не связаны со спецификой метода, они имеют общее значение.

Базы данных. В литературе описаны базы (банки) данных по характеристикам удерживания для отдельных вариантов хроматографического анализа; для различных НФ и ПФ; а также для разных условий разделения (см. обзоры в [232-234, 237). В БД могут включаться как значения абсолютных характеристик (например, абсолютные или исправленные времена удерживания), так и относительных (логарифмические или линейные индексы удерживания, относительные времена удерживания и др.), а также термодинамические характеристики (коэффициенты распределения). Кроме того, в БД включают информацию по другим свойствам эталонов, которые также могут быть поисковыми признаками. Например, коэффициенты чувствительности разных детекторов к одному и тому же X или отношения этих коэффициентов [235,238]. Индексы удерживания (в ГЖХ - индексы Ковача) в меньшей степени, чем абсолютные времена удерживания, зависят от температуры, скорости движения ПФ и количества НФ, поэтому большинство БД содержит в качестве поисковых признаков именно индексы. Относительные характеристики удерживания на обычной аппаратуре можно определить с погрешностью, не превышающей 1-3% [232], причем в реальной практике эта погрешность может быть значительно меньше. Поскольку процедура опознания по индексам Ковача без специального программного обеспечения является довольно длительной и трудоемкой, не потеряла своей актуальности и идентификация по абсолютным характеристикам. Абсолютные характеристики (например, время удерживания) воспроизводятся гораздо хуже, чем индексы Ковача, однако совершенствование аппаратуры, автоматизация и компьютеризация анализа позволяют существенно улучшить точность их определения.

В локальные БД включают данные по десяткам или сотням соединений, которые предположительно могут входить в состав проб определенного вида (например, известны

локальные БД для анализа наркотиков, пестицидов, эфирных масел, боевых отравляющих веществ, легких нефтепродуктов и др.); часто эти БД являются персональными. Соответствующие базы данных, как и соответствующие им информационно-поисковые системы, являются частью стандартного программного обеспечения хроматографов высшего класса, коммерческим продуктом. Существуют и глобальные БД, включающие информацию о характеристиках тысяч разнотипных соединений, не связанные с анализом объектов того или иного вида, а предназначенные прежде всего для исследовательских работ [234, 236, 239]. В начале 90-х годов считалось, что надежные данные по индексам удерживания получены примерно для 12 тысяч соединений. К сожалению, соответствующие данные разбросаны более чем по двум тысячам публикаций [240, 241].

Накопленная информация по удерживанию индивидуальных соединений позволяет вычислять характеристики тех же соединений в других условиях и, что важнее, прогнозировать характеристики неисследованных соединений [237, 240]. Значения t_X прогнозируются на основе известных корреляционных зависимостей, связывающих характеристики удерживания с числом атомов углерода в молекуле, температурой кипения, молекулярной массой или другими признаками соединений того же класса (лучше всего гомологов X), для которых значения t_X известны. Существование таких корреляций в принципе позволяет вводить в БД значения t_X лишь некоторых соединений данного ряда (как правило, не менее 4), а при поиске остальных использовать расчетные методы. Так как погрешность расчета может достигать 10-12 единиц индекса удерживания [230], в практике компьютерной идентификации такой способ используется относительно редко. Этот перспективный метод детально освещен, например, в многочисленных работах И.Г.Зенкевича [242].

Очень перспективной представляется также идея создания единого атласа хроматографических спектров, включающего систематизированные данные о величинах удерживания исследуемых веществ фазами разной полярности, а также о коэффициентах чувствительности детекторов. Попытки создания электронного атласа хроматографических спектров предпринимались несколько раз [237], но пока что достаточно полных атласов такого типа в распоряжении аналитиков нет, а неполнота БД может приводить к существенным ошибкам анализа.

3.2. **Компьютерные ИПС в хроматографическом анализе**

Базы данных по характеристикам удерживания и другим поисковым признакам обеспечивают работу компьютерных информационно-поисковых систем (ИПС) для

качественного хроматографического анализа реальных объектов. Такие ИПС появились еще в 70-е годы, в настоящее время они получили широкое распространение и даже стали объектом патентования [243]. "Хроматографические" ИПС особенно эффективны для решения задач групповой идентификации [244, 245] и для предварительного опознания индивидуальных компонентов соответствующих смесей. Групповая идентификация может проводиться по хроматографическому спектру, в том числе методами распознавания образов, а также другими способами, например, с помощью селективных детекторов. Групповая идентификация и даже оценка числа атомов углерода в молекуле X могут быть проведены и тогда, когда в БД нет табличных значений характеристик удерживания X [244]. Предварительная групповая идентификация существенно повышает надежность последующего опознания индивидуальных соединений, так как устраняет часть возможных межэталонных наложений. ИПС для групповой идентификации требуют отдельного обсуждения.

Возможность компьютерного качественного анализа конкретного объекта на уровне индивидуальных соединений (вплоть до распознавания изомеров) определяется наличием, полнотой и точностью локальной базы данных по характеристикам удерживания предполагаемых компонентов пробы, а также природой и сложностью объекта анализа. Отношение диапазона возможных значений ИУ к удвоенному значению стандартного отклонения ИУ (то есть разрешающая способность) в ГЖХ составляет 100-300 единиц для неполярных НФ и не превышает 100 единиц для полярных [246], именно этот параметр ограничивает сложность состава анализируемых смесей. Примерно к таким же ограничениям можно подойти, учитывая необходимость полного разрешения всех пиков исследуемой смеси и реальную эффективность лучших хроматографических колонок (порядка 10^5 тарелок).

Таблица 3.1.

Некоторые ИПС для хроматографической идентификации индивидуальных органических веществ

Метод	Объект анализа	Объем БД	Поисковый признак	Ссылка, год
ГЖХ	Растворители, краски	60	Отн. времена удерживания	1983 [247]
ВЭЖХ +УФ	Смеси пестицидов	74	Отн. удерживаемые объемы, отн. светопоглощение	1986 [248]

ГЖХ	Любые смеси	2,5 10^4	Индексы удерживания	1986 [249]
ГЖХ	Прямогонные бензины	4 10^2	Индексы удерживания	1993 [250]
ВЭЖХ	Лекарства, наркотики	до 10^3	Индексы удерживания, УФ-спектры	1994 [251]
ГЖХ	Воды и гидробионты (определение пестицидов)	до 10^2	Индексы удерживания	1997 [252]
ГЖХ	каталитические бензины	n 10^2	Индексы удерживания, отношения сигналов 2 детекторов	1999 [235]

В компьютерном качественном анализе пока не решена проблема расхождения табличных характеристик удерживания (даже относительных) с экспериментальными данными, получаемыми в конкретных лабораториях, как правило, такие расхождения статистически значимы, и для данной лаборатории их можно рассматривать как своеобразную систематическую погрешность эксперимента. Межлабораторная невоспроизводимость индексов удерживания (ИУ) одного и того же вещества (при использовании той же неподвижной фазы и аналогичных условий разделения) все еще слишком велика. В практике газожидкостной хроматографии эти расхождения иногда составляют 5-10 единиц ИУ для стандартных неполярных фаз и 15-25 единиц для полярных [237], тогда как при использовании капиллярных колонок различия в ИУ соседних хорошо разрешенных пиков могут составлять всего 3-4 единицы [253].

Межлабораторные расхождения связаны прежде всего с невозможностью точного воспроизведения свойств хроматографической колонки при ее повторном изготовлении (даже на основе одних и тех же материалов). Но полностью устранить “систематическую погрешность” характеристик удерживания не удастся даже при серийном выпуске идентичных (стандартных) колонок - по мере работы каждой колонки происходит индивидуальное изменение ее сорбционных свойств и постепенный дрейф характеристик удерживания, появляются систематические отличия t от t_x . Поэтому при эксплуатации ИПС аналитик должен проводить периодическую корректировку значений t_x по эталонам (модельным смесям). Это особо актуально, если применяются низкие значения d, обеспечивающие однозначную идентификацию.

Современные ИПС позволяют пользователю самостоятельно менять критерий d и даже устанавливать разные значения d для опознания разных веществ. Но выбор величины d представляет серьезную проблему: слишком жесткие требования к точности совпадений ($d \to 0$) могут привести к тому, что не будут опознаны действительные компоненты пробы, а слишком мягкие ($d >> 0$) приведут к "опознанию" отсутствующих веществ, то есть к ложным идентификациям [5, 254] Этот вопрос подробно рассмотрен в разделе 3.4.

Для наиболее простых объектов набор возможных компонентов априорно известен и ограничен. Если характеристики удерживания этих веществ не совпадают между собой и могут быть уточнены непосредственно в аналитической лаборатории, то надежность опознания существенно повышается, а результаты компьютерной идентификации принимают в качестве окончательных. Примером может быть качественный анализ бензинов по методике [250]. Аналогичные методики применяют и в анализе объектов окружающей среды [252]. Однако в случае смесей с очень сложным и непредсказуемым составом (что как раз соответствует экологическим объектам) пока что нельзя добиться желаемой надежности идентификации, пользуясь только характеристиками удерживания. Такая проверка целесообразна на стадии предварительного отбора "подозреваемых" - для отбраковки заведомо отсутствующих X. Проверка совпадения t с t_X должна дополняться проверкой по другим поисковым признакам и (или) исследованием спектров компонентов [227,231].

В методическом отношении опознание индивидуальных веществ по положению хроматографических пиков сходно с опознанием элементов по положению линий в спектре пробы; в обоих случаях применима одинаковая терминология и сходный математический аппарат [13]. Ранее для спектрального анализа были предложены вероятностные алгоритмы, позволяющие провести идентификацию с заданной надежностью [255]. Как уже отмечалось в разделе 3.1., при опознании хроматографических пиков, как и при отнесении спектральных линий, надо раздельно учитывать:

- а) случайные погрешности характеристик и число повторных испытаний;
- б) систематические расхождения табличных и реальных значений этих характеристик;
- в) межэталонные наложения.

Рассматривая фактор (а) и пользуясь методами теории вероятностей, можно априорно оценить надежность идентификации и выбрать (оптимизировать) значение критерия d. Влияние фактора (б) неоднократно изучалось [256, 257], но количественная связь межлабораторной невоспроизводимости справочных данных с надежностью идентификации пока не установлена. Влияние фактора (в) на достоверность компьютерной хроматографической идентификации не исследовано, однако известно, что вероятность

межэталонных наложений возрастает по гиперболическому закону с ростом числа компонентов смеси [258].

3.3. Вероятностная оценка надежности идентификации

Способы оценки надежности. В качестве простейшего способа часто рекомендуют вычислять *"информативность" анализа*, имея в виду отношение числа правильно отнесенных пиков на хроматограмме сложной смеси к общему числу пиков [230]. Не отрицая практической полезности подобных расчетов, мы полагаем, что их результат лишь косвенно связан с надежностью идентификации. К тому же сам термин "информативность" – в данном случае неудачен. Его лучше использовать для объективной оценки количества информации (по Шеннону), которое содержится в результатах соответствующего анализа [259].

Сомнительна и правомерность субъективного распространения полученных оценок на метод в целом, безотносительно к прибору, объекту анализа и кругу идентифицируемых веществ. Так, иногда указывают, что "информативность" качественного анализа по характеристикам удерживания (метод ГЖХ, одна колонка, неселективный детектор) составляет величину порядка 50-60% [231], хотя такое обобщение представляется слишком смелым.

Однако наиболее важно то, что любые обобщенные оценки (для всей хроматограммы) не характеризуют достоверность отнесения отдельного пика, надежность идентификации отдельного X (напрашивается сравнение с вероятностью выздоровления конкретного больного, которую нельзя оценить долей выздоравливающих по больнице в целом, как и средней температурой больных). Хорошо известно, что по хроматографическим данным одни компоненты пробы опознаются значительно легче и надежнее, чем другие. "Потребителя анализа" обычно интересуют лишь некоторые компоненты пробы (например, опасные токсиканты), и ошибки при обнаружении этих компонентов не могут компенсироваться правильным отнесением других пиков той же хроматограммы, даже при 99-процентной "информативности" анализа.

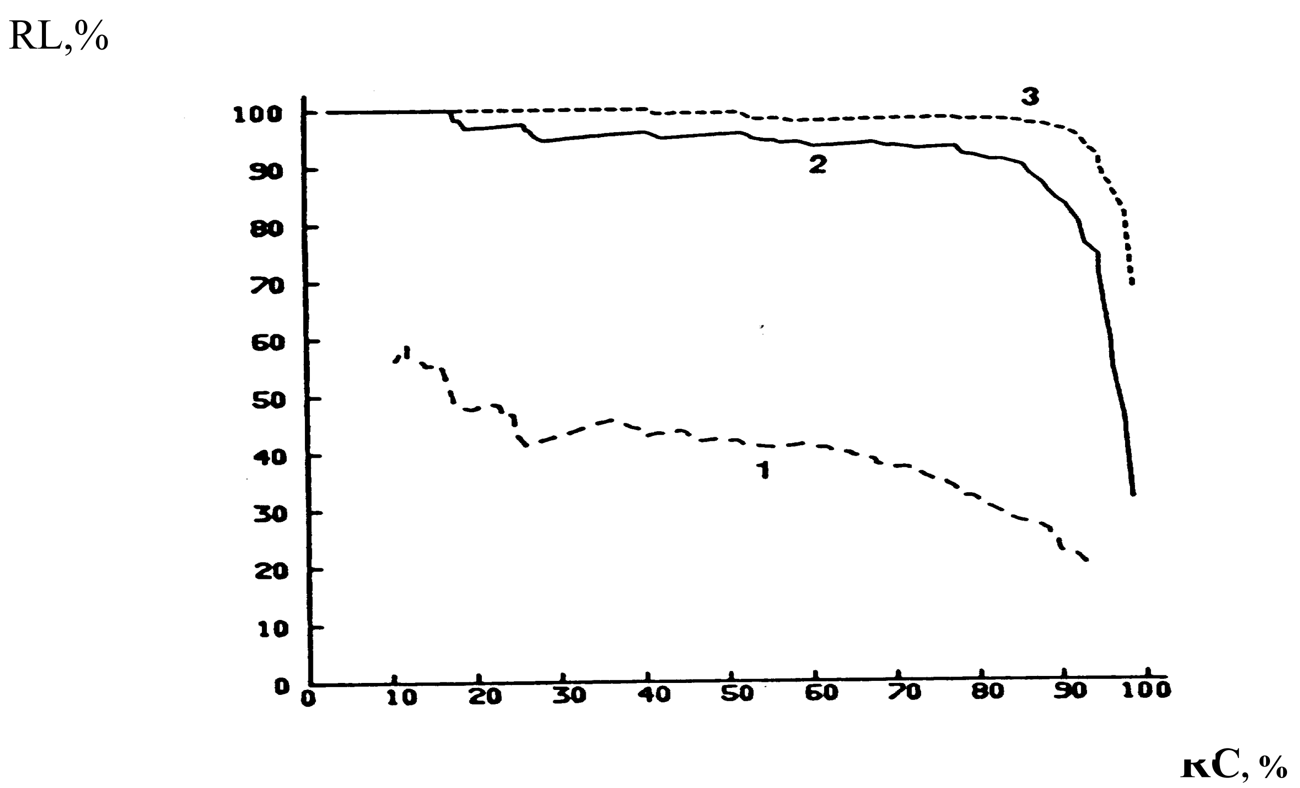


Рис.3.1. Оценка результатов идентификации пестицидов в модельных объектах по Мак-Лафферти. Расчеты проведены по характеристикам удерживания (1), по УФ-спектрам (2) и по комбинированным данным (3). Построено по данным [248]. Пояснения в тексте.

Результаты качественного анализа можно оценивать, пользуясь предложенными Мак-Лафферти [260] *идентификационными кривыми* (в англоязычной литературе по компьютерной масс-спектрометрии утвердился термин *recall / reliability plots, RC/RL*). Этот подход используют и в работах по хроматографической идентификации [248]. Алгоритм оценки не зависит от метода, объекта, поисковых признаков и конкретных критериев идентификации. Используют большой массив результатов компьютерного качественного анализа объектов с известным составом. По абсциссе RC откладывают отношение числа верно опознанных веществ к общему числу опознававшихся компонентов, по ординате RL – отношение числа верно опознанных веществ к общему числу опознанных (в том числе и ложно-идентифицированных). Использование менее строгих критериев опознания повышает результативность поиска, вероятность опознания *X* растет, но при этом появляются ложные идентификации, однозначность результатов анализа снижается. Обработку массива статистических данных проводят при разных критериях идентификации, каждый режим дает точку на кривой RL = f (RC). Полученные кривые весьма наглядны (рис.3.1). Видно, что применяя для опознания компонентов пробы только характеристики удерживания (кривая 1),

авторы работы [248] не всегда могли правильно опознавать присутствующие в пробе фосфорорганические пестициды (RC < 100%), зато в список опознанных всегда попадали и отсутствующие в пробе соединения (RL<<100%). Спектральная иденти-фикация компонентов после их разделения (кривая 2) давала значительно лучшие результаты, а комбинированный метод (ВЭЖХ+УФ) при определенных критериях (зона перегиба кривой) приводил к правильному и однозначному опознанию почти всех компонентов пробы.

К сожалению, оценка результатов поиска по методу Мак-Лафферти, как и оценка информативности по Ю.А.Другову, не дают возможности объективно, точно и метрологически обоснованно оценивать, а тем более сравнивать надежность идентификации отдельных компонентов пробы. Кроме того, такие оценки требуют наличия образцов с точно известным качественным составом; дают неверные результаты, если в пробе есть компоненты, чьи характеристики отсутствуют в БД; наконец, эти оценки невозможны без накопления очень большого объема экспериментальных данных.

Актуальной проблемой является априорная оценка надежности хроматографической идентификации отдельных компонентов исследуемой смеси [256]. Для случая спектральной идентификации аналогичные задачи решаются с учетом степени сходства спектров пробы и эталона, например, путем подсчета числа спектральных совпадений и последующего применения теории вероятностей [10]. Для хроматографического же анализа, где часто используется единственный поисковый признак, соответствующие алгоритмы неприменимы. Однако вероятностные оценки надежности идентификации возможны и в хроматографии. Соответствующие алгоритмы предложены в работах [254,261]. В основу этих алгоритмов положена идея раздельного расчета вероятностей двух различных ошибок, которые можно совершить, принимая решение о присутствии некоторого соединения в пробе:

1. Отсутствующий в пробе X после расшифровки хроматограммы может быть признан присутствующим (ложная идентификация, false positive). Вероятность этой ошибки обозначим символом α. Учет содержания БД и состава пробы для оценки α обязателен.

2. Присутствующий в пробе компонент X может быть признан отсутствующим. Назовем эту ошибку “пропуском сигнала” (false negative). Вероятность пропуска сигнала при единичном испытании обозначим символом β. Учет содержания БД и предполагаемого состава пробы для оценки β необязателен, поэтому оценить β значительно проще, чем α.

Выбор модели и схема расчета. Априорная оценка α и β на основе теории вероятностей возможна только в рамках некоторой модели. Примем следующие ограничения и допущения:

а) пики хроматограммы имеют гауссовскую форму и хорошо разрешены;

б) характеристикой положения пика (t) служит время удерживания. В этом случае t можно считать нормально распределенной случайной величиной [262]. Ее стандартное отклонение (σ) оценивается в предварительном эксперименте;

в) величина σ приблизительно одинакова для всех пиков пробы, соответствующие дисперсии статистически однородны;

г) воспроизводимость характеристик удерживания не меняется при повторном хроматографировании пробы или при переходе к другим пробам. Это позволяет не различать термины “воспроизводимость” и “сходимость”;

д) концентрации всех компонентов пробы в конечном разбавлении выше пределов их обнаружения с помощью используемого детектора.

е) в БД есть несовпадающие между собой значения t_x для всех предполагаемых компонентов пробы, точно измеренные в тех же условиях, в которых ведется анализ. Систематические расхождения t и t_x отсутствуют, математическое ожидание величины t для любого X независимо от концентрации равно t_x. Очевидно, это ограничение - самое важное и наиболее жесткое.

Физический смысл данной модели весьма прост: считаем, что и пропуск сигнала, и ложные идентификации происходят из-за случайного сдвига пиков пробы, а без него не наблюдаются. Разумеется, это не отвечает истине, но на первом этапе такая модель представляется оправданной. В дальнейшем модель должна постепенно усложняться, что приведет к уточнению оценок надежности. Решение о совпадении пиков в каждом испытании принимаем при $|t - t_x| < d$. Критерий d считаем одинаковым для всех пиков хроматограммы и не зависящим от состава пробы. Если критерий выбирается отдельно для каждого пика с учетом возможного присутствия в пробе компонентов Y_i с близкими к t_x значениями характеристик (t_Y), то соответствующий переменный критерий обозначается как d_x.

С помощью функций Лапласа (Ф) оценим вероятность пропуска сигнала. Вероятность случайного выхода пика X за пределы (t_x - d, $t_x + d$) равна:

$\beta = 1 - 2\,\Phi(d/\sigma)$. (-1).

По этой формуле можно рассчитать числовые значения β для некоторых d, выраженных в единицах стандартного отклонения, в “сигмах”. По результатам подобных расчетов построена кривая ***1*** на рис.3.2. В первом приближении величина β одинакова для разных X, не зависит от наличия других пиков на хроматограмме пробы, но зависит от воспроизводимости t и от выбора критерия совпадений, т. е. $\beta = \beta\,(d, \sigma)$. Пропуск сигнала

может не произойти даже при сильном сдвиге пиков, если в интервал $(t_x - d,\ t_x + d)$ вместо пика X случайно попадет пик другого компонента пробы. Поэтому учет состава пробы (второе приближение) несколько снизит величину β. Уточненное значение β неодинаково для разных X и для одного X в разных пробах.

Теперь оценим α. Причиной ложной идентификации X считаем сдвиг пика другого компонента пробы (Y_i) в интервал $(t_x - d,\ t_x + d)$. Обозначим через α_i вероятность принять пик единичного Y_i за пик X. Величина α_i должна быть функцией Δ_i, d_x и σ, где $\Delta_i = \left| t_{Y_i} - t_x \right|$.

Значения Δ_i определяют селективность t_x в рамках БД (заведомо отсутствующие в пробе Y_i предварительно исключают из БД). Вероятность случайного попадания пика Y_i в интервал ($t_x \pm d$) находим по разности:

$$\alpha_i = \Phi\left(\frac{\Delta_i + d}{\sigma}\right) - \Phi\left(\frac{\Delta_i - d}{\sigma}\right) \approx 0{,}5 - \Phi\left(\frac{\Delta_i - d}{\sigma}\right) \qquad (3\text{-}2).$$

Значение α должно определяться суммой α_i по всем предполагаемым компонентам данной пробы, с учетом близости их характеристик к t_x. Будем учитывать только возможное присутствие Y_1 и Y_2 - ближайших соседей X в ранжированном ряду, содержащем табличные значения характеристик удерживания всех предполагаемых компонентов данной пробы. А именно, $t_{Y_1} < t_x < t_{Y_2}$. Пренебрежение влиянием “дальних соседей”, естественно, несколько занизит α:

$$\alpha = \Sigma\, \alpha_i \approx \alpha_1 + \alpha_2 \qquad (3\text{-}3).$$

После подстановки получаем:

$$\alpha \approx 1 - \Phi\left(\frac{\Delta_1 - d}{\sigma}\right) - \Phi\left(\frac{\Delta_2 - d}{\sigma}\right) \qquad (3\text{-}4).$$

Для однозначной идентификации X критерий d_x не должен превышать Δ_1 и Δ_2. По формуле (3-4) можно найти α для любого расположения t_{Y_1} и t_{Y_2} относительно t_x.

Пример 1. Относительное время удерживания некоторого пика, измеренное на хроматограмме пробы, равно 0,825. В базе данных имеются табличные значения времен удерживания предполагаемых компонентов данной пробы для тех же условий разделения: 0,805 (вещество А), 0,825 (Б) и 0,855 (С). Пик идентифицировали, как принадлежащий Б, при этом пользовались критерием d = 0,005. Оценить надежность опознания Б, если воспроизводимость положения пиков на хроматограмме пробы характеризуется стандартным отклонением σ = 0,010.

Решение. Расчетные вероятности ошибок находим по формулам (3-4, 3-1):

$\alpha = 1 - \Phi(1{,}5) - \Phi(2{,}5) \approx 0.07$; $\qquad \beta = 1 - 2\Phi(0{,}5) \approx 0{,}62$.

Очевидно, вероятность ложной идентификации Б в рамках сделанных допущений невелика. Однако вероятен случайный пропуск сигнала и неопознание Б. Величину *d* можно и нужно было увеличить, это существенно снизило бы β при незначительном росте α (см. пример 2, раздел 3.4).

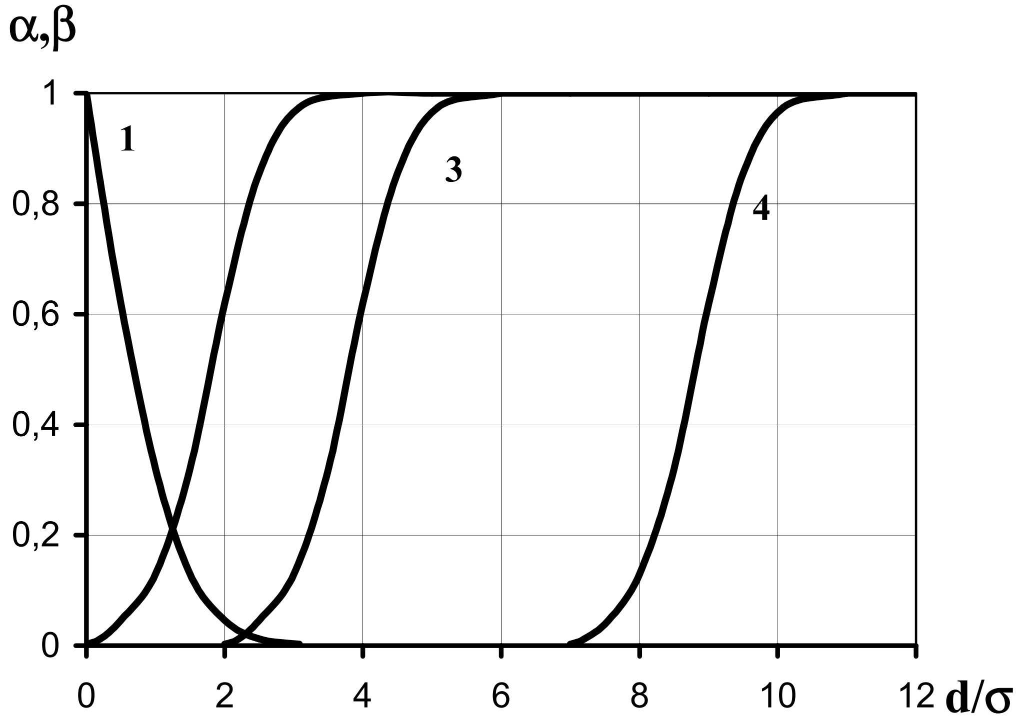


Рис.3.2. Расчетные вероятности ошибок идентификации в зависимости от критерия d при разной селективности табличных характеристик удерживания

Приведены значения β (кривая 1) и α (кривые 2-4) для $m = 3\sigma$ (2); 5σ (3) и 10σ (4).

При прочих равных условиях ложная идентификация *X* менее вероятна при симметричном расположении пиков "ближайших соседей" ($\Delta_1 = \Delta_2 = m$). В этом случае формула (3-4) упрощается, что позволяет найти минимальное значение α

$$\alpha_{min} \approx 1 - 2\,\Phi\left(\frac{m-d}{\sigma}\right) \qquad (3\text{-}5)$$

Формула (3-5) позволяет рассчитать вероятность ложной идентификации *X* при разных

значениях *d*, например для случаев, когда значения t_Y симметрично отстоят на 3, 5 и 10 “сигм” от t_x. (рис.3-1).

В работе [254] предложено оценивать суммарную надежность идентификации с помощью специальной функции:

$$P = 1 - \alpha - \beta \qquad (3\text{-}6),$$

подставляя в нее α и β, вычисленные по формулам (3-4) и (3-1). Эта функция может принимать значения от -1 до +1. Случай P =1 соответствует нулевой вероятности обеих возможных ошибок идентификации, а потому является идеальным (недостижимым на практике). В примере 1 величина P оказалась равной 1 - 0,07 - 0,62 = 0,31, что свидетельствует о невысокой надежности идентификации вещества Б.

Следует еще раз подчеркнуть, что выведенные формулы, учитывающие лишь один тип погрешностей, дают не истинные значения α и β, а лишь их нижние границы. Соответственно найденное значение P определяет не истинную надежность идентификации, а верхний предел этой величины.

Практическое использование вероятностных оценок. Анализ бензинов. Чтобы оценить реальную надежность хроматографической идентификации по характеристикам удерживания, рассмотрим конкретный пример - автоматизированное опознание индивидуальных углеводородов, входящих в состав прямогонных бензинов (детально эта проблема рассмотрена в статье [261]). Компьютерная идентификация в этом анализе предшествует расчету количественного содержания углеводородов и во многом определяет достоверность результатов в целом. Методика является стандартной для нефтеперерабатывающих предприятий [263]. Качественный состав бензинов определяли (до C_9 включительно) в режиме программирования температуры на хроматографе Perkin-Elmer Auto-System XL с капиллярной колонкой и системой автоматического ввода пробы. На хроматограммах имеется до 200 хорошо разрешенных пиков. Абсолютные времена удерживания фиксировались с точностью до 0,001 мин, а линейные индексы удерживания вычислялись с точностью до 0,1 единицы. Отнесение пиков вели параллельно с помощью двух “фирменных” ИПС: пакета программ *TurboChrom Navigator* и программы *Pianoeu* фирмы *Solutions.* Сопоставлялись времена (TurboChrom) или индексы удерживания (Pianoeu), критерии совпадений в обоих случаях задавались пользователем. Эталонные характеристики удерживания периодически уточнялись для данной колонки, прибора и режима разделения в ходе анализа бензинов с известным составом.

В рамках используемой модели надежность компьютерной идентификации является функцией воспроизводимости и селективности характеристик удерживания, а также зависит от выбора критерия идентификации. Рассмотрим эти факторы поочередно применительно к

определению углеводородного состава бензинов.

Воспроизводимость. Статистическая проверка показала, что при повторном вводе пробы абсолютные (а также относительные) времена удерживания варьируют приблизительно в соответствии с законом нормального распределения. Для индексов удерживания нормальное распределение наблюдалось лишь как исключение. Это соответствует данным [262]. Абсолютные времена удерживания воспроизводились на уровне коэффициента вариации (W), не превышающего 0,4%. Индексы удерживания воспроизводились точнее; для них $W < 0,1\%$; эти данные показывают, что современная хроматографическая аппаратура позволяет определять характеристики удерживания гораздо точнее и более воспроизводимо, чем считалось ранее [232]. Стандартные отклонения (S), выраженные в минутах, практически одинаковы для всех пиков, дисперсии статистически однородны, что и позволило вычислить величину σ ("сигму") для методики в целом.

Селективность. Для любой табличной характеристики t_x можно найти Δ_1 и Δ_2 - расстояния от пика *X* до соседних пиков (слева и справа) на идеальной хроматограмме, содержащей все компоненты локальной БД. Соседние углеводороды в ранжированной БД TurboChrom отличаются по временам удерживания в среднем на 0,6 минут, что с учетом найденной величины σ составляет около 10 "сигм". Малые ($\Delta < 3\sigma$) различия между соседними значениями t_x составляли примерно 20 % от общего объема БД. Точных совпадений по времени удерживания в этой БД нет, минимальное значение Δ составляет 1,5 σ. В более полной БД Pianoeu характеристики соседних веществ отличались меньше; после перевода ИУ во времена удерживания полученные значения различались в среднем на 5 "сигм". Малые различия наблюдались в 32 % случаев, а в 3 случаях времена удерживания двух разных углеводородов точно совпали. Для сравнения укажем, что селективность поисковых признаков в спектральном анализе в ряде случаев существенно хуже [264].

Разумеется, в хроматографии, как и в спектроскопии, селективность поисковых признаков не может рассматриваться изолированно от воспроизводимости измерений. Так, если использовать рассмотренные выше базы данных для расшифровки хроматограмм, полученных на другом приборе, с худшей воспроизводимостью (например, на ЛХМ-8МД σ почти на порядок выше), то все значения Δ при их выражении в "сигмах" оказались бы намного меньше, что указывает на снижение селективности). Компьютерная идентификация в этом случае станет менее надежной или вообще невозможной. Тот же эффект должен наблюдаться при переходе к качественному анализу проб более сложного состава, например, крекинг-бензинов [235]. Тогда в БД пришлось бы дополнительно включать все возможные компоненты соответствующих проб (например, непредельные углеводороды. практически отсутствующие в прямогонных бензинах). Это привело бы к уменьшению среднего

расстояния между значениями t_x и росту числа межэталонных наложений. Селективность поисковых признаков и надежность идентификации в этом случае также снижается, даже при использовании хроматографической аппаратуры с высокой воспроизводимостью измерения t.

Критерии совпадения пиков. Расчеты проводили с учетом используемых в каждой ИПС критериев совпадения. Критерий d при работе с программой *Pianoeu* одинаков для всех симметричных пиков и равен 0,5 (или, в некоторых случаях, 0,6) в шкале индексов Ковача. После перевода в единицы t этот критерий для разных пиков пробы оказывается несколько различным, но является величиной того же порядка, что и σ. В системе TurboChrom по умолчанию установлен критерий $d = 0{,}03\ t_x$, то есть вещество опознается, если различие между t и t_x не превышает 3%. В этой ИПС для всех компонентов пробы $d \gg \sigma$.

Таблица 3.2

Надежность опознания некоторых углеводородов в бензине (БД Pianoeu, *d* = 0,5I)

№ пика	*X*	*t*, мин	t_X, мин	Δ_1, мин	Δ_2, мин	α	β	*P*
6	2.2-диметилбутан	13.597	13.587	0.787	1.003	0	0.32	0.68
9	2-метилпентан	15.697	15.694	0.265	0.102	0.044	0.54	0.42
35	2,3-диметилгексан	36.393	36.409	0.780	0.175	0.014	0.23	0.76
49	1,2-диметилциклогексан	40.727	40.719	0.241	0.567	0.002	0.23	0.77
68	1,2,4-триметилциклогексан	47.312	47.305	0.373	0.290	0	0.21	0.79
76	3-метилоктан	49.637	49.613	0.184	0.284	0	0.27	0.73

Нули в столбцах вероятности означают, что α или β меньше 0.01.

Расчетные оценки надежности. Из табл.3.2 видно, что применение обычного режима работы ИПС Pianoeu с весьма строгим критерием d = 0,5 ИУ делает ложную идентификацию маловероятной (значения α не превышают 0,05), но вероятность неопознания присутствующих в пробе углеводородов довольно велика (β доходит до 0,6). Это указывает на неверный выбор критериев совпадения, величина d явно занижена. Для другой ИПС расчетная вероятность пропуска сигналов была исчезающе малой (для всех пиков $\beta < 0{,}001$),

зато вероятность ложных идентификаций была недопустимо большой (для большинства пиков $\alpha = 1$). Эти результаты также указывают на неудачный выбор критерия, величина d = 0,03 t слишком велика, что исключает возможность однозначного отнесения хроматографических пиков и допустимо только при последующем использовании дополнительных поисковых признаков.

Очевидно, в обоих случаях желаемая надежность качественного анализа бензина не достигается. Нужно оптимизировать критерии идентификации.

3.4. Выбор критериев идентификации

Приведенный в предыдущем разделе пример, связанный с анализом бензинов, показывает, что при компьютерной хроматографической идентификации критерий d должен устанавливаться с учетом воспроизводимости и селективности характеристик удерживания, однако обоснованных рекомендаций по этому вопросу в литературе нет. Разнобой мнений очень велик, так, авторы работы [256] для идентификации углеводородов по табличным значениям индексов удерживания рекомендуют значения d на уровне десятых долей ИУ, а авторы работы [257] – на уровне 5-10 единиц ИУ. Вероятностный подход дает более обоснованные рекомендации. А именно, при однократном хроматографировании пробы критерий d следует выбирать так, чтобы каждая из вероятностей α и β не превышала допустимый уровень (например, 0.05). По формуле (3-1) может быть выбрано минимальное значение d. Так, для $\beta < 0,05$ $d_{min} \approx 2\sigma$. Максимально допустимое значение d может быть найдено из формулы (3-5), если подставить в нее желаемое (пороговое) значение α и учесть селективность t_x. При этом d, Δ_1 и Δ_2 следует выражать в "сигмах". В частности, $\alpha < 0,05$, если $d = \Delta - 2\sigma$. Видно, что значения d_x должны отвечать условию:

$$2\sigma < d_x < \Delta - 2\sigma \qquad (3\text{-}7),$$

где Δ - меньшее из Δ_1 и Δ_2 для данного X. Выход за границы (3-7) приведет к случайным ошибкам идентификации, вероятность которых превысит 0,05. Если же применять значения d, находящиеся внутри найденных границ, например, использовать трехсигмовый критерий, то при высокой селективности характеристик удерживания можно пренебречь как возможностью пропуска сигнала, так и возможностью ложной идентификации X (α и β на уровне 0,01). Таким образом, для опознания хорошо разрешенных хроматографических пиков при высокой селективности характеристик удерживания целесообразно использовать единый критерий $d = 3\ \sigma$.

Неравенство (3-7) выполняется, если Δ больше 4σ. Если же это неравенство не

выполняется, то вероятность ошибочной идентификации будет больше 0,05 при любом d. Неравенство (3-7) важно не только для выбора критерия, но и для оценки аналитических возможностей метода в целом. Можно сформулировать следующее правило:

если в качестве поискового признака используются только характеристики удерживания, то хроматографическая идентификация веществ, для которых в базе данных не выполняется условие $\Delta/\sigma > 4$, принципиально ненадежна.

Независимо от выбранного критерия совпадения пиков, при невыполнении условия $\Delta/\sigma > 4$ надо применять дополнительные поисковые признаки (использовать селективные детекторы, исследовать спектры компонентов после их разделения и т.п.). Другой выход - многократно хроматографировать пробу, проверяя совпадения t с t_X на разных колонках или при разных режимах разделения. Если же указанное условие выполняется[++], надежная идентификация может быть достигнута и без дополнительных операций (табл.3.3). Разумеется, одновременно с проверкой условия $\Delta/\sigma > 4$ следует проверять выполнение исходных требований (полнота БД, отсутствие межэталонных наложений, отсутствие систематических погрешностей и т.д.).

Таблица 3.3

Расчетная надежность хроматографической идентификации

при разной селективности удерживания и разных критериях совпадения пиков

d/σ	**m = 2σ**	**m = 3σ**	**m = 5σ**	**m =10σ**
0,5	0,249	0,371	0,383	0,383
1	<u>0,367</u>	0,637	0,683	0,683
1,5	0,270	<u>0,733</u>	0,866	0,867
2,5	-	0,371	<u>0,975</u>	0,988
3	-	-	0,952	0,997
5	-	-	-	<u>1</u>
9	-	-	-	0,683
10	-	-	-	-

Примечание: в таблице приведены значения P для однократного хроматографирования пробы. Прочерк означает, что идентификация ненадежна ($P \leq 0$). Подчеркнуты максимальные значения P в каждом столбце, достигаемые при оптимальной величине d.

[++] При симметричном расположении данных ($\Delta_1 = \Delta_2 = m$) это условие записывается в форме $m/\sigma > 4$.

Отметим, что рекомендации того же типа ранее делались и ранее, но без теоретических обоснований. Так, авторы работы [253] считают, что для индексов удерживания внутрилабораторная воспроизводимость σ должна быть меньше 0,5 Δ.

Используя предложенную в работе [254] функцию *P*, можно оптимизировать критерий *d* даже при невыполнении указанного граничного условия. Несложно найти экстремум функции $P = f(d)$, приравняв нулю первую производную по аргументу *d*. Приводить математические выкладки нецелесообразно, но стандартные преобразования приводят к следующим выводам: а) при симметричном расположении характеристик удерживания Y_1 и Y_2 относительно t_x максимальная надежность идентификации достигается при $\alpha = \beta$, т. е. при равенстве вероятностей ошибок 1-го и 2-го рода; б) максимум функции $P = f(d)$ достигается при $d = 0,5$ m; в) надежность идентификации при оптимальном значении *d* определяется соотношением m/σ. Если пики на хроматограмме плохо воспроизводимы или табличные характеристики удерживания предполагаемых компонентов неселективны (в обоих случаях отношение m/σ снижается), кривые для α и β на рис.3-2 сблизятся и пересекутся намного выше оси абсцисс. В этом случае вероятности ошибок идентификации при любом *d* превысят допустимый предел.

Пример 2. Необходимо провести отнесение пика, у которого относительное время удерживания равно 0,825. В базе данных имеются табличные значения - 0,805 (А), 0,825 (Б) и 0,855 (С). Воспроизводимость положения всех пиков на хроматограмме пробы характеризуется стандартным отклонением $\sigma = 0,010$. Требуется выбрать оптимальное значение критерия *d* и оценить надежность идентификации Б.

Решение. В данном случае $2m = 0,050$, отношение $m/\sigma = 2,5$, граничное условие по селективности не выполняется, а поэтому критерий $d = 3\sigma$ применять не следует. Так как значения Δ_1 и Δ_2 различаются не очень сильно, считаем распределение характеристик удерживания приблизительно симметричным. Поэтому можно полагать, что надежность идентификации будет максимальна при использовании критерия $d_x = 0,5m = 0,0125$. В этом случае расчетные вероятности $\alpha = \beta = 1 - 2\Phi\ (d/\sigma) = = 1 - 2\Phi(1,25) \approx 0,21$. Тогда $P_{max} = 1 - \alpha - \beta = 0,58$. Более надежно опознать Б при однократном хроматографировании пробы и заданной воспроизводимости времен удерживания невозможно. (Учет несимметричности почти не изменит результат расчетов, уточненное значение $P_{max} = 0,52$). Таким образом, оптимизация критерия повышает надежность идентификации более, чем в два раза по сравнению с ранее использованным (пример 1) эмпирическим критерием.

Проиллюстрируем эффект оптимизации критерия на том же реальном примере,

что в предыдущем разделе. Опознание индивидуальных углеводородов в прямогонных бензинах по методике [263] можно вести, задавая разные значения d и подсчитывая число углеводородов, отвечающих условию $|t- t_X| < d$. Результаты для некоторых пиков на хроматограмме бензина показаны в табл. 3.4. Видно, что при d = 0,5% t "окно совпадений" оказалось слишком узким, в ряде случаев наблюдался пропуск сигнала. При d = 1,5% t и более широких окнах таких случаев не было.

С другой стороны, увеличение d приводит к ухудшению однозначности, один и тот же пик на хроматограмме пробы можно приписать разным углеводородам. Неоднозначность отнесения заметнее для более тяжелых компонентов пробы, так как абсолютная величина d (в минутах) для них больше. Эти эмпирические результаты соответствуют изменению расчетной вероятности возможных ошибок (пропуска сигналов и ложных идентификаций).

Таблица 3.4

Однозначность отнесения хроматографических пиков при компьютерном качественном анализе бензина. Влияние выбора критерия совпадений, по данным [261]

№ пика	t, мин	d =0.5% t	d=1.5%t	D=3.0%t	d=6.0%t	$d=3\sigma$
7	13,63	1	1	1	1	*1*
10	15,75	0	1	3	3	*1*
38	36,60	0	2	6	12	*1*
68	47,34	1	4	9	19	*1*
76	49,52	2	3	9	18	*2*

Указано число эталонов в БД TurboChrom, отвечающих условию $|t- t_X| < d$

Неоднозначность отнесения пиков на практике маскируется тем, что обычно "фирменные" ИПС из нескольких веществ, удовлетворяющих условию $|t- t_X| < d$, отбирают и выдают пользователю название того вещества, у которого модуль разности $(t - t_x)$ минимален (т.е. каждому пику пробы ставится в соответствие ближайший пик на хроматограмме стандартного бензина). Разумеется, при случайном варьировании характеристик удерживания нет гарантии, что такой пик будет принадлежать действительно присутствующему компоненту пробы. Именно поэтому в некоторых случаях повторный

качественный анализ одной и той же пробы бензина с применением системы TurboChrom давал сомнительные результаты; при том же режиме расшифровки один и тот же пик пробы при повторном ъроматографировании приписывался разным веществам. По-видимому, ИПС должны выдавать пользователю названия всех веществ, удовлетворяющих идентификационным критериям. Это позволит объективнее судить о работе ИПС и, что важнее, при неоднозначной идентификации проверять результат с помощью других поисковых признаков, в том числе с применением других методов анализа.

Данные табл. 3.4 показывают также, что критерий $d = 3\sigma$ обеспечивает не только снижение вероятности обеих идентификационных ошибок до пренебрежимо малого уровня, но и обеспечивает правильное и однозначное отнесение большинства хроматографических пиков (см. последнюю колонку). Трехсигмовый критерий не обеспечивает желаемых результатов (однозначности отнесения) только в тех случаях, когда для опознаваемого X величина Δ_1 или Δ_2 меньше 3σ. Примером может быть отнесение пика № 76 в табл.3.4 (предположительно, 3-метилоктан), поскольку эталонные характеристики удерживания 3-метилоктана при данной методике разделения углеводородов практически совпадают с характеристиками 3-этилгептана, идентификация неоднозначна при любых критериях. Именно в таких случаях следует привлекать дополнительные поисковые признаки, либо повторять анализ в других условиях.

И теоретические расчеты по формуле (3-4), и эмпирические данные (табл.3.4) свидетельствуют, что применяемые на практике постоянные эмпирические критерии совпадения пиков приводят к неодинаковой надежности опознания разных компонентов одной и той же пробы, так как их характеристики удерживания имеют различную селективность. Поэтому для одинаково надежной идентификации разных компонентов пробы в работе [265] рекомендуется использовать индивидуальные критерии $d_X = 0{,}5$ m, вычисляемые компьютером отдельно для каждого X. Однако экспериментального подтверждения эта идея пока еще не получила.

3.5. Хроматографическая идентификация при многократных испытаниият

Невозможность надежной идентификации индивидуальных соединений при проведении качественного анализа сложных объектов по характеристикам удерживания на одной колонке и при одном режиме разделения давно привели хроматографистов к использованию трех-четырех колонок с резко различными свойствами [11]. Надежность идентификации в таком варианте анализа обуславливается, в частности, рациональным выбором неподвижных фаз. Важно обеспечить реализацию разных типов межмолекулярных

взаимодействий фазы с разделяемыми соединениями (электронодонорные, дисперсионнные и электроноакцепторные эффекты) [266]. Используют наборы колонок с неполярной, среднеполярной и сильнополярной неподвижными фазами [227]. При этом важна эффективность всех колонок, обеспечивающая полное разделение компонентов. В некоторых случаях используют единственную колонку, но проверку повторяют при разных температурах в методе ГЖХ или разном составе элюента в методе ВЭЖХ [267]. Все эти варианты объединяет термин "*многомерная хроматография*". Считают опознанными те вещества, для которых совпадение t и t_x наблюдается на <u>каждой</u> из полученных хроматограмм. Один из наиболее интересных вариантов этого метода - многоступенчатая хроматография - предполагает последовательное соединение нескольких колонок и отвод элюата к детектору после каждой ступени [244]. Направлять в следующую колонку можно не весь элюат, а лишь часть его, определенную фракцию. В результате достигается лучшее разделение смеси, а сопоставление полученных хроматограмм позволяет, как и при параллельном соединении колонок, провести идентификацию более надежно за счет уменьшения вероятности межэталонных наложений.

Наиболее серьезной проблемой многомерной хроматографии при анализе сложных объектов является несоответствие порядка выхода компонентов на нескольких хроматограммах одной и той же пробы. Иногда с переходом от одной НФ к другой меняется и общее число пиков. Из-за этого очень трудно указать, какие пики на разных хроматограммах соответствуют одному и тому же *X*, что существенно облегчило бы и групповую, и индивидуальную идентификацию *X*. В этих случаях помогают несколько приемов: использование селективных детекторов, сопоставление сигналов двух детекторов (отношение их сигналов для пика *X* не зависит от типа колонки, но зависит от природы *X* [235]), предварительная химическая обработка пробы (вычитание или смещение пиков), а также многоступенчатое разделение пробы.

Повторные испытания обеспечивают возможность групповой идентификации. Так, разность ИУ на двух колонках разной полярности приблизительно одинакова у соединений одного гомологического ряда, но сильно меняется с переходом к веществам других классов [11], в качестве группового поискового признака можно использовать и температурный градиент, и некоторые термодинамические параметры.

Метод многомерной хроматографии пока не нашел широкого применения при компьютерной идентификации органических соединений, отчасти это связано с необходимостью иметь несколько БД по характеристикам удерживания, а отчасти с невозможностью передать в машинных алгоритмах опыт и интуицию опытного хроматографиста, которые позволяют ему соотнести пики на разных хроматограммах без

формальных критериев, а иногда - и без табличных значений t_x. Тем не менее несомненно, что развитие метода многомерной хроматографии также приведет к компьютеризации анализа, и следует заранее указать на опасность, ожидающую будущих разработчиков и пользователей соответствующих ИПС. А именно, в случае применения формальных критериев совпадения пиков вероятность ошибочных результатов в некоторых случаях может возрастать по мере увеличения числа повторных испытаний. Это положение представляется неожиданным, не отмечавшимся ранее в литературе и весьма важным с практической точки зрения, поэтому оно требует более детального обсуждения.

Вероятность ошибок при повторных испытаниях. Допустим, что ввиду плохой воспроизводимости характеристик удерживания требуемая надежность идентификации некоторого *X* не достигается даже при оптимальном значении критерия d_x. Для достижения требуемой надежности проще всего *n* раз повторить ввод пробы (одна и та же колонка, один и тот же режим работы хроматографа), а затем проверить совпадение усредненной характеристики удерживания ($\bar{t}$) с табличным значением t_x [256]. В этом случае удается выявить и исключить грубые промахи. Еще важнее то, что переход к новой (в $\sqrt{n}$ раз меньшей) величине σ не изменит оптимального значения d, но уменьшит дисперсию случайной величины $\bar{t}$ в n раз по сравнению с дисперсией t , соответственно уменьшит значения α и β. Следовательно, надежность идентификации, характеризуемая по формуле (3-6) функцией *P,* повысится.

Однако на практике проверяют не только совпадение усредненной величины *t* с t_X (с точностью до *d*), но и попадание *каждого* значения *t* в интервал ($t_x \pm d$). Такое требование ведет к повышению вероятности пропуска сигнала, необнаружению реально присутствующего *X*, и это может быть подтверждено расчетом. Пусть единичные вероятности α и β при повторных испытаниях постоянны. Исходы испытаний рассматриваем как независимые события. Вероятность пропуска пика *X* хотя бы в одном из *n* проведенных испытаний (β_n), как и вероятность его ложной идентификации во всех испытаниях (α_n), можно рассчитать по несложным формулам, представляющим частный случай формул Бернулли :

$$\alpha_n = \alpha^n \qquad (3\text{-}8),$$

$$\beta_n = 1 - (1 - \beta)^n \qquad (3\text{-}9).$$

В табл.3.5 приведены значения α_n и β_n, вычисленные по (3-8) и (3-9) для разных *n*. Наблюдаемое падение α_n с ростом *n* легко объяснимо: маловероятно, чтобы каждый раз происходила бы случайная ложная идентификация одного и того же *X*, вероятность этого события по мере роста n будет стремиться к нулю. С другой стороны, с ростом *n*

увеличивается вероятность выхода пика X за пределы “окна совпадений” хотя бы в одном из n испытаний, а это - при формализации алгоритма - должно приводить к необнаружению присутствующего X. Таким образом, при проведении повторных испытаний в одинаковых условиях следует проверять совпадение только по усредненным, а не по единичным значениям характеристик удерживания.

Таблица 3.5.

Изменение расчетной вероятности идентификационных ошибок при проведении повторных испытаний

n	α_n			β_n		
1	0,5	0,2	0,05	0,20	0,10	0,05
2	0,25	0,04	0,0025	0,488	0,190	0,098
3	0,125	0,008	-	0,590	0,271	0,142
4	0,031	-	-	0,672	0,410	0,226
5	-	-	-	0,893	0,651	0,401

Единичные вероятности указаны в строке $n = 1$ *и предполагаются неизменными. Прочерк означает, что* α_n *или* β_n *менее 0,001.*

Теперь рассмотрим проведение повторных испытаний *в разных условиях* (варьирование типа колонки, температуры разделения, состава элюента и др.). В этом случае значения t усреднять нельзя, каждое значение t надо сопоставлять с соответствующим ему по условиям разделения табличным значением t_x , причем значения $t_X = t_X(i)$ различны. В каждом (i-ом) испытании проверяют принадлежность пика $t=t(i)$ интервалу $\{(t_X(i) - d, t_X(i) + d\}$. Требование, чтобы при *каждом* испытании пик X попадал в табличный интервал, приводит к недопустимо высоким значениям вероятности пропуска сигнала. Если единичная вероятность пропуска сигнала во всех испытаниях остается одной и той же, то расчет можно провести по формуле (3-9), в других случаях расчетные формулы усложняются, но сохраняется та же тенденция - даже при не очень больших n величина β_n быстро достигает 1. Задавая максимально допустимое значение β_n , можно рассчитать, какое наибольшее число испытаний можно провести, не опасаясь неопознания X за счет случайного сдвига пика хотя бы в одном случае из n. После логарифмирования (3-9) получаем:.

$$n_{max} = \frac{\lg(1-\beta_n)}{\lg(1-\beta)} \qquad (3\text{-}10)$$

В качестве примера рассмотрим влияние числа повторных испытаний для случая, когда α = 0,5, а β = 0,05. Желательно путем увеличения числа испытаний снизить вероятность ложной идентификации хотя бы до уровня α_n < 0,05. Расчет по формуле (3-8) показывает, что это возможно при n = 5, но вероятность пропуска сигнала в одном из испытаний и, соответственно, неопознания присутствующего X в соответствии с формулой (3-9) достигнет 0,40.

Таким образом, при не очень хорошей воспроизводимости t опасно проводить большое число повторных проверок в разных условиях! По-видимому, как и при выборе критерия d, при оценке необходимого числа повторных испытаний должен быть достигнут компромисс, причем желаемый уровень α_n определит нижнюю, а β_n - верхнюю границу оптимизируемого параметра n – числа повторных испытаний в одинаковых условиях.

Чтобы достичь одновременного снижения вероятности обеих идентификационных ошибок, целесообразно ввести понятие *критического числа совпадений* (аналогичная идея давно реализована в спектральном анализе), а критическое число совпадений можно рассчитать методами теории вероятностей.

Будем считать достаточным, чтобы попадание t в интервал ($t_x \pm d$) происходило бы в k или более случаев из n проведенных испытаний. Вероятность ложной идентификации X не менее чем в k случаях из n испытаний обозначим как $\alpha_n(k)$, а вероятность пропуска X более, чем в (n-k) случаях - как $\beta_n(k)$. т выбора критического значения k будет зависеть вероятность обеих ошибок идентификации. При постоянных единичных вероятностях α и β в каждом (i-ом) испытании значения суммарных вероятностей можно вычислять по формулам Бернулли:

$$\beta_n(k) = \sum_{i=0}^{k-1} C_n^i \beta^{n-i}(1-\beta)^i = 1 - \sum_{i=k}^{n} C_n^i \beta^{n-i}(1-\beta)^i, \qquad (3\text{-}11),$$

$$\alpha_n(k) = \sum_{i=k}^{n} C_n^i \alpha^i (1-\alpha)^{n-i}, \qquad (3\text{-}12),$$

где C_n^i число сочетаний из n по i.

Если задан допустимый уровень α_0 (например, $\alpha_0 = 0,.05$), то нужно подобрать, исходя из формулы (3-12), самое малое k при котором $\alpha_n(k) < \alpha_0$, а затем рассчитать по формуле (3-11) $\beta_n(k)$. *Для* подбора величины k и вычислений вероятностей $\beta_n(k)$ и $\alpha_n(k)$ можно использовать специальное программное обеспечение [268] (без него вычисления слишком

трудоемки). Математические обоснования метода и алгоритм подбора критического числа совпадений даны в работе [269]. Результаты серии подобных расчетов приведены в табл. 3.6.

Видно, что величина *k* может быть существенно меньше *n*, что весьма важно в практическом отношении, и при этом вероятности обеих ошибок идентификации останутся в требуемых пределах. При $n = 2$ или *3* уменьшения *k* по сравнению с *n* обычно не происходит.

Таблица 3.6.

Подбор критического числа совпадений для разных начальных условий

α	β	**n**	α_0	**k**	$\alpha_n(k)$	$\beta_n(k)$
0.1	0.1	2	0.01(0.05)	2	0.01	0.19
0.1	0.1	3	0.01	3	0.001	0.27
0.1	**0.1**	**5**	**0.01(0.05)**	**3**	**0.0086**	**0.0086**
0.1	**0.1**	**10**	**0.05**	**4**	**0.013**	$\mathbf{9.1 \cdot 10^{-6}}$
0.1	0.2	4	0.01/0.05	3	0.004	0.05
0.1	**0.2**	**7**	**0.01**	**4**	**0.0027**	**0.0027**
0.1	**0.2**	**7**	**0.05**	**3**	**0.026**	$\mathbf{1.8 \cdot 10^{-4}}$
0.2	0.2	3	0.01(0.05)	3	0.008	0.488
0.2	0.2	5	0.01(0.05)	4	0.0067	0.2627
0.2	0.2	7	0.01	5	0.0047	0.148
0.2	**0.2**	**7**	**0.05**	**4**	**0.0333**	**0.0333**

Жирным шрифтом выделены результаты, одновременно удовлетворяющие следующим условиям: $\alpha_n(k) < \alpha_0$; $\beta_n(k) < 0,05$; $k < n$.

Путем последовательных вычислений при разных *n* легко решить задачу нахождения минимального число повторных испытаний n_0, которое при заданных α и β дает требуемую достоверность идентификации. Естественно, n_0 увеличивается по мере роста α и β. Оптимизация функции $P = 1 - \alpha_n(k) - \beta_n(k)$ может быть достигнута путем машинного перебора 2–3 близких *k*. Применяя программное обеспечение [268], можно быстро (за 3-5

минут) подобрать минимальное количество повторных испытаний и критическое число совпадений.

С другой стороны, оптимизация значений k и n дает возможность уменьшить вероятности обеих идентификационных ошибок одновременно. Так, если взять (как в приведенном выше примере) α = 0,5, а β = 0,01, но задаться уровнем k = 2 при n = 3, то расчетные вероятности ошибок окажутся менее 0,01, а именно: $\beta_3(2)$=0.00725, $\alpha_3(2)$=0.00125. Оптимальное число совпадений (одновременно обеспечивающее снижение $\alpha_n(k)$, и $\beta_n(k)$ до безопасного уровня) тем больше, чем выше вероятности α и β в единичных испытаниях.

Применение описанных выше алгоритмов требуется предварительных исследований воспроизводимости и селективности удерживания *X* в разных условиях. Эти алгоритмы могут обеспечить заданную надежность компьютерной идентификации только в рамках модели, описанной в разделе 3.3. Разумеется, надежность идентификации может быть повышена и эмпирическим путем, вне рамок той или иной модели - например, за счет применения “идентифицирующих” детекторов [270], в том числе масс-спектрометрического или ИК-спектрометрического опознания каждого выходящего из колонки компонента.

Глава 4
КАЧЕСТВЕННЫЙ СПЕКТРАЛЬНЫЙ АНАЛИЗ НЕРАЗДЕЛЕННЫХ СМЕСЕЙ

4.1. Общие подходы к качественному анализу смесей

Несмотря на огромные успехи хроматографического анализа, далеко не все природные и техногенные смеси могут быть быстро и полностью разделены до индивидуальных соединений. Если такое разделение и достигается, то, как показано в гл.3, использование хроматографических характеристик удерживания не всегда ведет к надежной идентификации компонентов смеси. Необходима проверка по дополнительным признакам [231], а она требует либо весьма сложной и дорогостоящей аппаратуры, либо больших затрат времени. Поэтому интерес к качественному спектральному анализу неразделенных проб сохранился и после выявления всех возможностей ГЖХ и ВЭЖХ. Появление же компьютеров усилило интерес к этому научному направлению, особенно перспективному для скрининга трудноразделяемых микропримесей.

Информационно-поисковые системы, созданные для опознания чистых веществ и описанные в главе 1, не могут использоваться для анализа сложных смесей. Проявление

пиков посторонних веществ снижает степень совпадения спектров пробы и ее i-го компонента, а то и ведет к признанию i-го компонента отсутствующим. Как правило, поиск дает удовлетворительные результаты лишь для основного компонента смеси. Поиск следующего ведут после вычитания из спектра смеси спектра опознанного соединения. В 70-е годы этот прием применяли довольно часто, но, ввиду накопления погрешностей, - не более чем для 3-4 последовательно обнаруживаемых компонентов [271]. Альтернативным способом был расчет условных концентраций всех предполагаемых компонентов пробы по алгоритмам, родственным методу Фирордта, при этом использовали заранее определенные коэффициенты [272]. Компоненты, у которых условные концентрации оказывались выше некоторого критерия, считали идентифицированными.

Однако к концу 80-х годов выяснилось, что более перспективны совершенно другие варианты компьютерного качественного анализа смесей. А именно: а) предварительная кластеризация спектра пробы и б) непосредственный поиск признаков i-го компонента в спектре пробы. Соответственно, идеи, лежащие в основе этих вариантов, можно условно назвать *методологией выделения субспектров* и *методологией обратного поиска*.

Первый подход основан на выделении из спектра смеси некоторых кластеров - совокупностей спектральных признаков, относящихся к отдельным компонентам пробы. Например, выделение ряда пиков, интенсивность которых согласованно меняется при перераспределении концентраций смеси или варьировании условий регистрации спектра. Для выделения таких кластеров (субспектров) предложено множество приемов и математических алгоритмов [273]. Компоненты смеси можно опознавать по субспектрам таким же образом, как и индивидуальные вещества, в частности, с помощью обычных ИПС прямого поиска. Таким образом, выделение субспектров сводит идентификационную задачу к уже рассмотренным в данной книге (гл.1 и 2). Чаще всего выделение и отождествление субспектров используют в масс-спектрометрии.

Обычно для выделения субспектров последовательно требуется:

- получить каким-либо способом из исходной пробы набор *родственных смесей* (то есть смесей с одинаковым качественным, но различным количественным составом);
- зарегистрировать в идентичных условиях спектры этих смесей;
- обработать все полученные спектры на ЭВМ.

Родственные смеси получают при неполном хроматографическом или термодиффузионном разделении компонентов; часто пробу просто вводят в источник масс-спектрометра. Можно обойтись одной смесью (исходной пробой), но в таком случае ее спектры придется снимать несколько раз, меняя условия так, чтобы добиться различного

вклада компонентов. Число полученных "родственных" спектров должно быть не менее числа компонентов данной смеси, и переопределенность системы повышает точность выделения субспектров [274]. Задача упрощается, если число реально присутствующих компонентов заранее известно.

Для выделения субспектров созданы многочисленные программные продукты, реализующие такие алгоритмы, как метод Аленцева-Фока, разные варианты факторного анализа, метод варьирования концентраций, метод сопряженных градиентов, корреляционный метод и др. [274, 275]. Как показано в работах М.С.Хоца, условиями применимости и эффективности всех этих алгоритмов являются:

- строгая аддитивность спектров компонентов;
- достоверное различие спектров всех компонентов;
- неизменность формы спектральной кривой компонента при изменении состава смеси или условий регистрации;
- прямопропорциональная зависимость аналитического сигнала от концентрации компонента;
- минимизация уровня фона, шумов и т.п.

Но даже при выполнении всех этих условий однозначного и устойчивого решения часто не удается добиться без введения дополнительных допущений о форме спектральных кривых, наличии участков, свободных от межэталонных наложений и т.п. В случае сравнительно простых смесей (2-3 структурно несходных компонента) выделение субспектров проходит весьма точно, что и позволяет вести компьютерную идентификацию [276]. Иногда субспектры вполне удовлетворительно опознаются даже в случае их выделения из спектра сложной смеси (например, шестикомпонентной). При усложнении состава смеси надежность идентификации ухудшается, так как становится трудно провести различие между *субспектрами*, отвечающими действительно присутствующим соединениям, и *псевдоспектрами*, возникающими вследствие ошибок регистрации, нестабильности спектральных параметров и неточностей расчета [277]. Для дифференциации таких спектров предлагались различные критерии, как эмпирические, так и метрологически обоснованные. Введение последних позволяет надежно установить масс-спектрометрическим методом состав некоторых весьма сложных смесей без разделения [278].

Очевидно, развитию данного аналитического метода и его применению в рутинном анализе препятствуют два фактора - а) невыполнение условий, обеспечивающих точное выделение субспектров (аддитивность, отсутствие фона и т.п.); б) необходимость дополнительных операций получения родственных смесей или родственных спектров.

Второй подход, основанный на методологии обратного поиска, реализуется в ИПС, непосредственно сопоставляющих эталонные спектры всех предполагаемых компонентов пробы с ее реальным спектром [10]. Возможность присутствия компонента количественно оценивается, исходя из степени проявления его спектральных признаков в спектре смеси (см. раздел 1.2.2). Соединения, у которых эта характеристика окажется выше критического уровня, считаются опознанными. Эта методология прослеживается в публикациях многих авторов, и далее ей будет уделено особое внимание. Как и методология выделения субспектров, она не связана с тем или иным типом спектров или определенным классом объектов, ее с успехом можно применить в разных методах анализа (табл.4.1): в люминесцентном анализе при комнатной температуре (ЛЮМ), в спектрофотометрическом анализе в УФ-области, в ИК-спектроскопии и в других методах, где расшифровываемые спектры имеют непрерывный характер. Наиболее целесообразен такой подход при расшифровке линейчатых спектров, то есть в рентгенофазовом анализе (РФА), в масс-спектрометрии (МС), в атомно-эмиссионном спектральном анализе (АЭм). Обратный поиск используют и при расшифровке квазилинейчатых спектров низкотемпературной люминесценции (НЛ), гамма-спектров радионуклидов (γ), спектров рентгеновской флуоресценции (РФЛ).

Отметим сразу же, что ни тот, ни другой из указанных подходов к анализу неразделенных смесей нельзя рассматривать как “основной” и тем более “единственно правильный”.

Таблица 4.1.

Примеры ИПС для компьютерной идентификации веществ по спектрам смесей

Год, название	**Метод**	**N-БД**	**Особенности**	**m_x**
1968 JV	РФА	$2{,}5\cdot10^4$	прямой поиск, эмп.критерии	2-3
1974 PBM	МС	10^2	обратный поиск, учет характеристичности	3
1976 PBM*	МС	$1{,}8\cdot10^4$	постоянные вероятностные критерии	3
1979 Фазан	РФА	$5\cdot10^2$	расчет условных концентраций, эмп. критерии	3
1982 АРФА	РФА	10^4	построение идентифицирующих функций (ИФ)	3
1983 Руслан	МС	?	учет информативности признаков	4-6

1983 АКФА	РФА	$2{,}5 \cdot 10^3$	Расчет условных концентраций, эмп. критерии	4-5
1984 FXQUAL	ЛЮМ	?	Диалоговый режим, интуитивная идентификация	3
1985 PBM*	МС	$7{,}6 \cdot 10^4$	Вычитание субспектров, добавочный прямой поиск	3-4
1985 Спектр-1	НЛ	10^2	учет характеристичности, бинарные коды, ИФ	3
1986 PAWMI	ИК	62	Вычитание субспектров, структурные корреляции	4
1987 Фазан*	РФА	$4 \cdot 10^4$	Расчет условных концентраций, эмп. критерии	10
1987 Спектр-2	НЛ	10^2	Переменные вероятностные критерии	6
1989 Спектр-3	НЛ	200	Учет $I_{отн}$, вероятностные критерии	15
1993 Аргус	АЭм	70	переменные вероятностные критерии	6
1994 Sampo 90	γ	$2{,}6 \cdot 10^3$	Одновременное использование 50 разных признаков	32

Обозначения: m_x - максимальное число компонентов, идентифицируемых по одному спектру. *N -БД* - число соединений в базе данных*;* *эмп.критерии* - постоянные для всех компонентов эмпирические критерии*;* * - новая версия. Более подробная сводка с литературными ссылками приведена в работе [10].

Методологии выделения субспектров и обратного поиска дополняют друг друга, имея и немало общего. Так, в обоих случаях требуется применение БД и ИПС. Анализ возможен лишь там, где компоненты смеси не взаимодействуют друг с другом и, более того, присутствие одного компонента не влияет на спектральные признаки другого (аддитивные смеси). В обоих случаях надежность идентификации тем выше, чем менее похожи друг на друга эталонные спектры компонентов и чем многочисленнее их поисковые признаки. Оба подхода не являются универсальными, они применимы не для всех смесей и не для всех видов спектров.

Области эффективного применения рассматриваемых подходов не совпадают. Так, серьезным ограничением методологии обратного поиска является невозможность опознания компонентов, эталонные спектры которых не включены в БД. Поэтому рассмотренные в главе 2 мощные алгоритмы выявления структуры неизвестного соединения неприменимы для интерпретации спектра смеси, но вполне могут использоваться в отношении субспектров. Дополнительным преимуществом выделения субспектров является одновременная оценка и качественного, и количественного состава пробы. Этот метод

применим даже в анализе нефтепродуктов [279], но, разумеется, не для идентификации и количественного определения каждого из сотен присутствующих в пробе индивидуальных углеводородов, а для решения задач структурно-группового анализа.

С другой стороны, известные алгоритмы выделения субспектров далеко не всегда обеспечивают точность, требуемую для успешной работы ИПС; надежность идентификации веществ по субспектрам оставляет желать лучшего. Обратный же поиск позволяет быстро и без дополнительных операций получать надежные результаты даже в таких сложных случаях, как опознание индивидуальных токсикантов в атмосферном воздухе или в сточных водах нефтеперерабатывающих предприятий [280].

4.2. Методология обратного поиска в анализе смесей

Качественный спектральный анализ пробы сложного состава требует детального исследования спектра пробы - последовательной проверки присутствия всех возможных компонентов данной пробы с использованием соответствующих эталонов. Разумеется, пользователь ИПС не работает с реальными эталонными веществами. Фактографические БД содержат информацию, полученную ранее по сотням тысяч таких веществ, например, их эталонные спектры. Проверка включает серию испытаний (сопоставлений) по ряду заранее отобранных и включенных в БД поисковых признаков каждого предполагаемого компонента. Естественно, разные признаки имеют неодинаковую значимость для опознания этого соединения (разную характеристичность). Цель проверки - поиск признаков, совпадающих у пробы и эталона с точностью до погрешности измерения, и регистрация таких совпадений. Естественно, достоверность идентификации тем выше, чем большая доля поисковых признаков совпадает у пробы и эталона, и чем более характеристичным (или даже специфическим для проверяемого компонента) является каждый из них..

В разделе 3.1. уже были даны определения некоторых терминов, необходимых для понимания алгоритмов поиска (компонент, признак присутствия, испытание, проверка). Дополним этот перечень.

Характеристичность *признака (weight, uniqueness)* - статистический вес, с которым j-ый признак i-го эталона должен учитываться при обнаружении компонента X_i в данной пробе; далее обозначается символом G_{ij}. Так, иногда считают, что характеристичность линии определяется ее относительной интенсивностью в эталонном спектре X_i . С другой стороны, считают, что для достоверного обнаружения X_i наиболее важно проявление тех признаков, которые в пределах БД характерны только для X_i . Компьютерный анализ БД позволяет выявить признаки, специфические для X_i , определить значения G_{ij} для других -

неспецифических - признаков, а затем так рассчитать характеристичность каждого признака, чтобы значение G_{ij} падало от 1 до 0 при переходе от специфических к общим признакам. Величина G_{ij} должна зависеть от R - числа предполагаемых компонентов пробы, имеющих данный признак, а также от M – общего числа сопоставлявшихся эталонных спектров, Так, в работе [264] при $R > 1$ используется функция $G_{ij} = M / (M+200) (R-1)$, причем при $R = 1$ величина G_{ij} принимается равной единице. Например, если в БД есть 85 эталонных спектров предполагаемых компонентов пробы ($M = 85$), а некоторая спектральная линия присутствует (с учетом $\Delta\lambda$ - погрешности измерения длины волны) в трех эталонных спектрах ($R = 3$), то X_{ij} для данного признака равно 0,1492. Линия, общая для 10 соединений, при опознании каждого из них с помощью той же БД должна учитываться в гораздо меньшей степени ($G_{ij} = 0{,}033$), и т.д.

При таком подходе значения G_{ij} оказываются не постоянными, а зависящими от состава пробы и точности измерений. При использовании ЭВМ значения характеристичностей нетрудно каждый раз рассчитывать заново (см. раздел 4.5). Функции характеристичности можно конструировать так, чтобы учитывалась не только специфичность каждого признака, но и относительная интенсивность линии (пика) или другие особенности признаков.

Сигнал присутствия i-го компонента (*qualitative signal,* S_i) - количественная оценка результатов проверки. В простейшем случае $S_i = n_i$, то есть сигнал равен числу характерных для X_i признаков, проявившихся при проверке данной пробы. Более точно можно рассчитать сигнал по формуле (4-1):

$$S_i = \sum_{1}^{n_i} G_{ij} \qquad (4\text{-}1).$$

Сигнал присутствия (суммарная характеристичность проявившихся признаков X_i) несет информацию о составе пробы; зависит (нелинейно) от концентрации X_i, а также от условий испытаний; при повторных проверках варьирует из-за случайных погрешностей; в отсутствие X_i может принимать ненулевые значения (фон, шум) из-за неспецифичности признаков. Таким образом, предложенный для качественного анализа в работе [13] термин "сигнал присутствия" близок понятию "аналитический сигнал" в количественном анализе. Эти аналогии отражают методологическое единство качественного и количественного анализа. Вместе с тем сигнал присутствия имеет свою специфику; это результат расчета, безразмерная величина, ее не следует смешивать с традиционным аналитическим сигналом - результатом прямого измерения.

Чем выше сигнал присутствия опознанного компонента, тем выше достоверность идентификации. Однако нельзя сопоставлять значения S_i у разных компонентов одной

пробы, так как они изначально обладают неодинаковым количеством признаков [13, 272, 281]. Интерпретация сигнала присутствия должна зависеть от l_i, поэтому S_i нормируют. Например, в работе [113] нормированный сигнал (A_i) рассчитывали так:

$$A_i = S_i / S_{i\,max} = \sum_{1}^{j=n_i} G_{ij} \;/\; \sum_{1}^{j=l_i} G_{ij} \qquad (4\text{-}2).$$

Величина A_i принимает любые значения от 0 до 1, в зависимости от того, какая часть признаков X_i проявилась при проверке данной пробы, и какова их характеристичность. По A_i объективно сопоставляют сигналы присутствия разных компонентов, оценивают влияние посторонних веществ, оптимизируют условия регистрации [282]. Возможны и другие, более сложные способы оценки сигнала присутствия: по соотношению "сигнал/шум" или по относительной информативности [283]. Частным случаем нормированного сигнала присутствия при опознании чистых веществ является *степень совпадения (Z)*. Способы оценки степени совпадения уже обсуждались в гл.1.

Критерий идентификации (*criterion)* - заданное пользователем критическое значение сигнала присутствия в его обычной (S_i), нормированной (A_i) или неявной (n_i, Z) форме. Компонент признается идентифицированным, если его сигнал присутствия превосходит соответствующий критерий [6, 284 и др.]. Основной теоретической проблемой при создании ИПС для анализа смесей является обоснованный выбор критерия идентификации. Именно этот выбор определяет вероятность ошибок, а тем самым и достоверность результата анализа.

Заметим, что сопоставление сигналов присутствия с некоторыми критериями характерно и для традиционных методов качественного анализа, не связанных с применением компьютеров, но в этих методах и сигналы присутствия, и критерии оцениваются интуитивно. Так, считают необходимым проявление в спектре пробы <u>нескольких</u> линий, совпадающих по своему положению с линиями отыскиваемого элемента [285]. Аналогичные рассуждения можно найти в литературе по дробному качественному анализу ионов [286] (проверка по нескольким реакциям) или в руководствах по хроматографическому анализу. В последнем случае требуют совпадения времен удерживания с табличными значениями на нескольких разных колонках. К сожалению, во многих публикациях, связанных с качественным анализом, критерии идентификации либо вообще не указываются, либо о них говорится весьма неопределенно. Так, сообщается о совпадении нескольких пиков по длине волны, но не указывается, при каких значениях $\Delta\lambda$ пики считали совпадающими и почему выбрано именно это значение $\Delta\lambda$. Бытующие в литературе утверждения типа: “...определение расположения нескольких наиболее

интенсивных линий каждого элемента вполне достаточно для того, чтобы *вне всякого сомнения* утверждать, что данный элемент присутствует в пробе" [285] — без уточнения понятия "несколько", без указания типа спектра и т.п. некорректны. Достоверность спектральной идентификации, в том числе и по количеству совпадающих линий, нуждается в строгих количественных оценках.

Размытость и произвольность обычных критериев идентификации приводят к тому, что естественные вопросы типа: "*Вы идентифицировали переданный на анализ белый порошок как героин, а какова вероятность ошибки*?" - пока не находят ответа. Это вызывает резкие и обоснованные возражения метрологов [5]. Очевидно, в любом методе качественного анализа, основанном на проверке наличия признаков X_i, признание компонента опознанным должно сопровождаться оценкой неопределенности идентификации, подобно тому, как результат количественного анализа сопровождается оценкой его погрешности. В обоих случаях выводы имеют вероятностный характер и должны сопровождаться количественными оценками.

4.3. Алгоритмы анализа смесей с применением вероятностных критериев

Алгоритмы ИПС, созданных для опознания чистых веществ и для анализа сложных смесей, имеют много общего. В обоих случаях проводится предварительная отбраковка части эталонов и образуется сокращенная рабочая библиотека (РБ), для каждого оставшегося в РБ эталона рассчитывается некоторая количественная характеристика, а ранжирование оставшихся эталонов по этой характеристике используется для формирования машинного ответа. Однако вышеуказанные количественные характеристики, как и способ формирования машинного ответа - совершенно различны.

При опознании чистых веществ обычно оценивают степень совпадения спектров пробы и эталона по относительной интенсивности линий ($I_{отн}$) на всех длинах волн. В анализе же смесей применение $I_{отн}$ как поискового признака нецелесообразно, так как значение $I_{отн}$ линии в спектре смеси может резко отличаться от $I_{отн}$ той же линии в i-ом эталонном спектре - из-за межэталонных наложений, эффекта внутреннего фильтра и по другим причинам. О присутствии i-го компонента судят по числу линий i-го эталонного спектра, совпадающих по длине волны (частоте, массовому числу и т. п.) с линиями в спектре пробы; при этом учитывают погрешность измерений ($\pm\Delta\lambda$), условия возбуждения спектров, возможность межэталонных совпадений и другие факторы.

Другое отличие связано с организацией отбора компонентов. При опознании чистых веществ пользователь обычно получает список компонентов, занявших после ранжирования

N первых мест, и сам принимает решение, какой компонент считать опознанным. В принципе такой подход применим и в анализе простейших смесей.

Зависимость сигнала присутствия компонента от его порядкового номера в ранжированном списке иногда позволяет дифференцировать действительно присутствующие компоненты пробы от остальных, так как на соответствующей кривой (“*идентифицирующей функции*”) наблюдается излом [6]. Однако этот излом сглаживается в результате попадания на первые места структурных аналогов действительно присутствующих компонентов, а также из-за случайных совпадений (рис.4.1). Поэтому применение идентифицирующих функций лишь в самых простых случаях позволяет точно определить состав смеси [287, 288] и не может считаться универсальным способом дифференциации отсутствующих и присутствующих компонентов.

Альтернативный способ организации машинного ответа - выдача списка компонентов, у которых сигнал присутствия в данной пробе оказался выше некоторого критерия. Этот способ в анализе смесей является основным. Естественно, тот же прием можно использовать и при опознании чистых веществ, но в соответствующих ИПС критический уровень степени совпадения постулируют довольно редко. Обычно в число опознанных попадают присутствующие в пробе соединения (не обязательно все!) и некоторые отсутствующие (*псевдокомпоненты).* Последним термином мы обозначаем опознанные компьютером вещества, которые в исследуемой пробе заведомо отсутствуют. Как правило, псевдокомпоненты - структурные аналоги действительно присутствующих компонентов смеси, но ложные идентификации могут быть и следствием случайного совпадения признаков; в этом случае структура псевдокомпонентов может быть совершенно несходной со структурой соединений, действительно присутствующих в пробе.

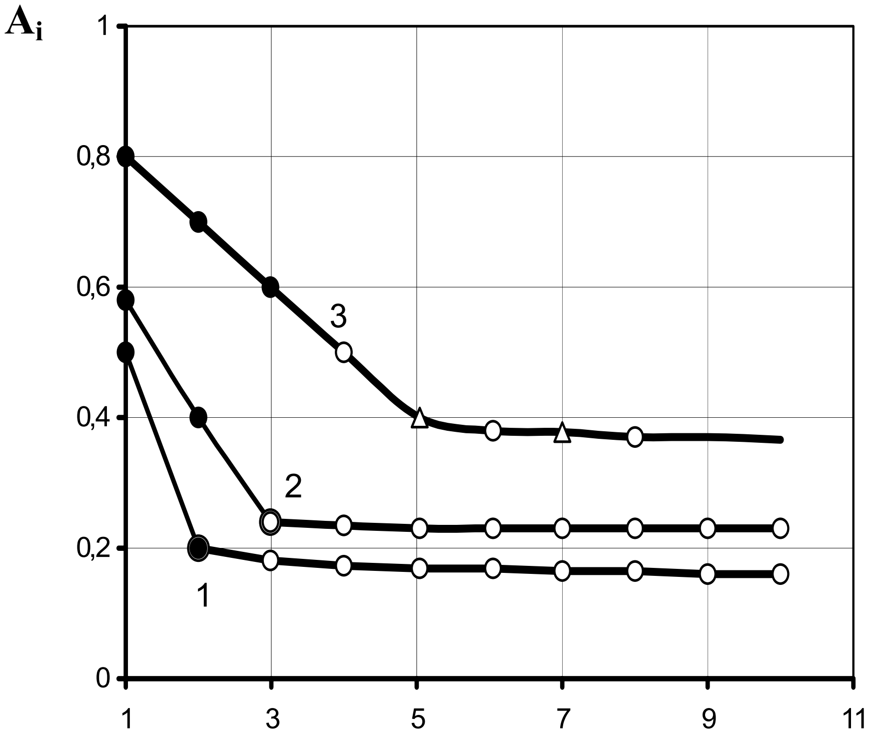


Рис. 4.1.

Идентифицирующие функции при опознании полиаренов по спектрам НЛ

Примечание. Исследуемые пробы - растворы индивидуального полиарена (1), бинарной смеси (2), четырехкомпонентной смеси (3). Сигналы присутствующих компонентов показаны черными кружками, их структурных аналогов - треугольниками, прочих соединений - светлыми кружками. Методика регистрации спектров описана в [282], методика опознания – в [288].

Обычно применяют эмпирические (произвольные) критерии (примером может быть работа [284]). Реже критерии определяются пользователем на основании работы с обучающими выборками. Но и в том, и в другом случае на практике пользуются постоянными критериями, одними и теми же для всех проверяемых компонентов, хотя несомненно, что численные значения критерия должны зависеть от природы компонента, от состава пробы и от условий регистрации спектра. В идеальном случае они должны заново рассчитываться компьютером при проверке каждого X_i и при расшифровке каждой новой пробы [13, 94]. Такие расчеты возможны на основе теории информации [104, 283] или теории вероятностей [6, 137].

Предложено использовать в качестве критерия величину $n_{i\ крит}$ - максимальное число случайных совпадений, которое с вероятностью α может наблюдаться в отсутствие компонента X_i и его структурных аналогов [289]. Если в ходе проверки число реальных совпадений между спектром пробы и эталонным спектром X_i (далее это число обозначается

как n_i) превысит величину $n_{i\ крит}$, то можно сделать предварительный вывод о присутствии X_i в пробе. Наоборот, при $n_i \leq n_{i\ крит}$ совпадения могут быть и случайными, решить вопрос о присутствии X_i без привлечения дополнительных признаков нельзя. Для окончательного вывода о наличии X_i в пробе надо (даже при $n_i >> n_{i\ крит}$) учитывать относительную интенсивность линий эталонного спектра, их характеристичность именно для X_i, условия возбуждения и регистрации спектра пробы и другие дополнительные признаки. Теоретический расчет $n_{i\ крит}$ (или других вероятностных критериев) представляет весьма сложную задачу.

Обоснованный выбор $n_{i\ крит}$ возможен либо на основании теоремы Бернулли, либо с помощью многократного моделирования спектра пробы и машинной расшифровки полученных псевдоспектров. Второй вариант близок известному в хемометрике бутстрэп-методу [290]. Он позволяет дифференцировать отсутствующие и присутствующие компоненты более точно, чем с помощью критериев, найденных расчетным методом, но требует больших затрат машинного времени и в меньшей степени апробирован на практике.

Рассмотрим способы приближенной оценки вероятностных критериев, основанные на теореме Бернулли. В теории спектральной идентификации этот подход был предложен в работе [137], алгоритмы расчета критериев уточнены в статье [255], а их программная реализация и практическое использование описаны, например, в работах [94,113].

В целом вероятностный подход к решению идентификационных задач основывается на признании результата единичного испытания случайным событием. Как и любой результат измерения, результат испытания по j-ому признаку i-го предполагаемого компонента пробы отягощен случайными погрешностями. В частности, межэталонные наложения спектральных линий, ошибки при измерении их положения, появление ложных линий, связанных с погрешностями регистрации, и другие факторы могут привести к тому, что результат единичного испытания с некоторой вероятностью P_{ij} будет положительным и в в отсутствие X_i (или его спектрального аналога). И наоборот, по разным причинам с некоторой вероятностью $^*P_{ij}$ результат испытания может быть отрицательным и в присутствии X_i.

После проведения серии испытаний число ошибочных положительных результатов может превысить заданный критический уровень $n_{i\ крит}$ (далее для удобства записи используем более краткое обозначение - n_{kp}). Это приведет к ложной идентификации X_i. Вероятность (α) такого события можно рассчитать, если знать все P_{ij}. Аналогичным образом, зная все $^*P_{ij}$, можно рассчитать β - вероятность случайного пропуска реально присутствующего в пробе компонента X_i. Очевидно, теория идентификации должна включать расчет вероятности обеих случайных ошибок (α и β по отдельности) при любом

$n_{кр}$. В этом случае будет возможен компромиссный выбор критерия, при котором вероятность каждой из ошибок (ложной идентификации и пропуска сигнала) не превысит допустимого уровня, и только в этом случае можно будет опознавать компоненты с заданной надежностью. В гл.3 показано, как можно получить компромиссное решение аналогичной задачи в хроматографическом анализе. Однако в спектральном анализе расчет компромиссных критериев значительно труднее: не ясно, как оценить значения $^*P_{ij}$, а следовательно и вероятности неопознания действительно присутствующих компонентов пробы. Рассмотрим поэтому вопрос о вероятностных критериях без учета такой возможности, добиваясь лишь предотвращения ложных идентификаций.

Примем в качестве первого приближения, что единичные вероятности случайного совпадения с отсутствующим X_i (значения P_{ij}) одинаковы для всех признаков X_i и, более того, для всех предполагаемых компонентов данной пробы, тогда $P_{ij} \equiv P$. Примем также, что линии сопоставляемых спектров пробы и эталона равномерно распределены по всему интервалу (λ_1, λ_2), в котором проводится сопоставление линий, то есть случайные спектральные совпадения с линиями отсутствующих X_i равновероятны для всех линий пробы. Совпадающими будем считать линии пробы и эталона, если их положение в шкале длин волн отличается не более, чем на $2\Delta\lambda$, где $\Delta\lambda$ - максимальная погрешность измерения положения линии. Разобьем мысленно (λ_1 , λ_2), на m спектральных окон шириной $2\Delta\lambda$. Проявление каждой линии в спектре пробы можно, пользуясь терминами теории вероятностей, рассматривать как независимое испытание, а попадание линии в определенное спектральное окно — как один из равновероятных исходов. Положительный исход (совпадение с одной из линий X_i) произойдет, если линия пробы окажется в одном из l_i, спектральных окон, сформированных вокруг линий компонента X_i . Тогда единичная вероятность случайного совпадения линий равна:

$$P = \frac{l_i}{m} = \frac{2\Delta\Delta \cdot l_i}{|\lambda_2 - \lambda_1|} \qquad (4\text{-}3).$$

Спектр пробы, содержащий N линий, можно рассматривать как серию из N независимых испытаний. Вероятность одновременного положительного исхода n независимых испытаний из N проведенных рассчитывается по формуле Бернулли.

$$P_{N,n} = \frac{N!\,P^n(1-P)^{N-n}}{n!(N-n)!} \qquad (4\text{-}4).$$

Формула (4-4) позволяет рассчитать вероятность случайного совпадения спектров пробы и эталона по n линиям. С ее помощью можно рассчитать и вероятность того, что в серии испытаний число положительных исходов превысит значение n, равное $n_{кр}$:

$$\alpha \ = \ P(\,n > n_{кр}\,) \ = 1 - \ \sum_{0}^{n_{кр}} P_{N,n} \qquad (4\text{-}5).$$

Очевидно, α - вероятность случайного совпадения пробы с отсутствующим компонентом по числу линий, большему, чем $n_{кр}$. Если в ходе анализа решение о неслучайном характере совпадений и следовательно, обнаружении X_i - принимается в случае, когда $n > n_{кр}$, то α — уровень значимости нуль-гипотезы о случайном характере совпадений и отсутствии X_i .

Прямой расчет α по формулам (4-4) и (4-5) при больших N весьма трудоемок. Его можно облегчить, аппроксимируя зависимость $P_{N,n}$ от n формулами нормального распределения:

$$\alpha \approx 1 - \int_{-\infty}^{t_{кр}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} d\text{t}, \qquad (4\text{-}6).$$

Значения интеграла Гаусса (далее $Г(t)$) в формуле (4-6) для любых t хорошо известны, а величина $t_{кр}$ получается в результате нормировки:

$$t_{кр} = \frac{n_{кр} - NP}{\sqrt{NP(1-P)}} \qquad (4\text{-}7).$$

Для практики важна обратная задача: нахождение $n_{кр}$ для заранее заданного значения α, например 0,05 или 0,01. Из (4-7) следует:

$$n_{кр} = NP + t_{кр}\sqrt{NP(1-P)} \qquad (4\text{-}8)$$

Подстановка (4-3) в (4-8) и приводит к искомому значению $n_{кр}$. Табличная величина $t_{кр}$ находится из условия $Г(t) = 1 - \alpha$, например, для $\alpha = 0{,}05$ $t_{кр}=1{,}65$. Чтобы уменьшить вероятность ложных идентификаций, рассчитанные по формуле (4-8) дробные значения $n_{кр}$ надо округлять до целочисленных значений в большую сторону. Полученные значения критерия - приблизительные, но в большинстве случаев они точно совпадают с результатами прямого расчета по формуле (4-5). Естественно, для разных условий регистрации спектра пробы, для разных проб, для разных компонентов одной пробы значения $n_{кр}$ будут различными.

Пример расчета. В спектре пробы (например, квазилинейчатом спектре низкотемпературной люминесценции) содержится 40, а в эталонном спектре некоторого вещества - 20 линий; оба спектра сняты в интервале 300-500 нм с погрешностью 0,2 нм. Какое число спектральных совпадений требуется для того, чтобы считать вещество присутствующим в пробе с надежностью 0,95?

Решение. По формуле (4-3) $P = 0{,}04$. Требуемый уровень значимости нуль-гипотезы -

0,05 определяет, что $t_{кр}$=1,65. Следовательно, $n_{кр}$ = 40 0,04 + 1.65 √ 40 0,04 0,96 = 3.64 ≈ 4 . Расчет по формуле (4-5) приводит к той же величине. Следовательно, случайным характером одновременного совпадения пяти, шести или большего числа линий пробы с линиями X_i можно пренебречь, такие совпадения с 95 %-й надежностью доказывают присутствие вещества X_i (или его аналога со сходным набором признаков). Если же в ходе проверки будут найдены лишь 4 (или менее) линий X_i, то случайным характером этого события пренебречь нельзя, компонент не может считаться идентифицированным с требуемой надежностью, и в результате проверки ИПС должна его отбраковать.

В табл. 4.2 и 4.3 приведены значения $n_{кр}$, рассчитанные по формуле (4-8) для некоторых других комбинаций N, l_i, $|\lambda_2 - \lambda_1|$, α и $\Delta\lambda$. Из табл. 4.2 следует, что снижение точности измерений при прочих равных условиях повышает вероятность случайного проявления линий любого предполагаемого компонента пробы, в том числе отсутствующего. Поэтому, чтобы избежать ложных идентификаций, в таких случаях необходимо требовать совпадения по большему числу линий. Расчеты показывают, что иногда совпадения 1-2 линий пробы и эталона с точностью до 0,001 нм вполне достаточно для надежной идентификации. Это соответствует обычной практике атомно-эмиссионного элементного анализа при использовании спектрометров с высокой дисперсией. Однако совпадения по 1-2 линиям недостаточно для опознания, если положение линий измеряется с точностью до 0,1 нм или еще хуже, что соответствует применению приборов с низкой дисперсией или сильному уширению спектральных линий. Для обоснованного вывода в этих случаях требуется значительно большее число совпадений; тем большее, чем выше требуемая надежность идентификации $P_{иден} = 1 - \alpha$. Так, при расшифровке квазилинейчатых спектров люминесценции надежная идентификация полиаренов требует, как правило, совпадения по 5-10 линиям.

Необходимо вновь подчеркнуть, что при опознании разных компонентов по одному и тому же спектру пробы значения $n_{кр}$ не могут быть одинаковы (из-за различия в l_i). При идентификации одного и того же компонента в разных пробах $n_{кр}$ также меняется от пробы к пробе, поскольку меняется N, а возможно - и другие характеристики спектра. Заданные же раз и навсегда значения $n_{кр}$ получить в принципе невозможно.

Таблица 4.2

Расчет критериев для некоторых случаев спектральной идентификации: влияние точности измерений (*$\Delta\lambda$*) и желаемой надежности идентификации *$P_{иден}$* = *1-* α

$\Delta\lambda$	$P_{иден}$	P	$n_{кр}$
0,01	*0,95*	*0,002*	*1*
0,05	*0,95*	*0,010*	*2*
0,20	*0,95*	*0,040*	*4*
0,20	*0,99*	*0,040*	*5*
0,20	*0,999*	*0,040*	*6*
1,0	*0,999*	*0,20*	*13*

Примечание*: N = 40; l_i =20;* $|\lambda_2 - \lambda_1|$ *= 200 нм.*

Таблица 4.3

Расчет критериев для некоторых случаев спектральной идентификации: влияние заселенности спектров

N	l_i	$\lvert\lambda_2 - \lambda_1\rvert$, нм	*P*	$N_{кр}$
10	*10*	*200*	*0,02*	*1*
40	*10*	*200*	*0,02*	*3*
40	*40*	*200*	*0,08*	*6*
40	*40*	*50*	*0,32*	*18*
40	*40*	*25*	*0,64*	*29*
100	*40*	*25*	*0,64*	*72*

Примечание: $\alpha = 0,05$; $\Delta\lambda = 0,2$ *нм*

Из табл. 4.3 следует, что при постоянной погрешности измерений положения линий величина $n_{кр}$ возрастает при увеличении *заселенности* сопоставляемых спектров, то есть при увеличении числа линий, приходящихся в спектрах пробы или эталона на одно “спектральное окно”. При высоких заселенностях, когда $N \approx m$ или $l_i \approx m$, величина $n_{кр}$ может превысить l_i (см. последнюю строку в табл.4.3). В таких случаях совпадения даже всех линий эталонного спектра с линиями пробы будет недостаточно для идентификации соответствующего вещества с требуемой надежностью! Ясно, что качественный спектральный анализ с подсчетом совпадений имеет, как и любой метод, ограниченную применимость: при высокой заселенности сопоставляемых спектров и, соответственно, высокой вероятности случайных совпадений - описанными выше вероятностными критериями пользоваться не следует.

В рассмотренной выше упрощенной модели [137] учитывались лишь основные факторы, определяющие величину $n_{кр}$: число линий в сопоставляемых спектрах и погрешность измерения длины волны. Расчет $n_{кр}$ вели приближенным способом, а не по точной формуле (4-5). В модели не был учтен и объем рабочей библиотеки (число последовательно проверяемых эталонов). Однако основной недостаток рассмотренного выше алгоритма – принятие неочевидного постулата об одинаковой вероятности всех случайных совпадений и, следовательно, одинаковой значимости любого совпадения для общего результата проверки. Разная вероятность случайных совпадений для разных линий спектра может быть следствием неравноточности измерений, неравномерного размещения спектральных линий по рассматриваемому интервалу длин волн, а также других факторов [255]. Проблема еще более осложняется ввиду закономерно меняющейся по (λ_1, λ_2) вероятности межэталонных наложений. Влияние всех этих факторов частично можно учесть, если находить критерии в компьютерном эксперименте, путем многократной генерации и расшифровки псевдоспектров (см. 4.4).

Необходимо лишь уточнить, что независимо от способа расчета вероятностные критерии в принципе не могут предотвратить идентификацию соединений, отсутствующих в пробе, если те являются структурными (а следовательно и спектральными) аналогами действительно присутствующих компонентов. Применение вероятностных критериев – необходимое, но не достаточное условие получения точных результатов.

4.4. Применение вероятностных критериев при расшифровке спектров низкотемпературной люминесценции. Оценка критериев в ходе компьютерных экспериментов

Экспериментальную проверку вышеизложенного метода (в принципе приложимого к спектрам любого типа) проводили при расшифровке бинарно кодированных квазилинейчатых спектров низкотемпературной люминесценции[2]. В качестве модельного объекта для оптимизации стратегии обратного поиска использовали смеси полициклических ароматических углеводородов (полиарены, ПАУ). Спектры НЛ полиаренов и их смесей в определенных условиях имеют квазилинейчатый характер (эффект Шпольского). Положение квазилиний в шкале длин волн определяется с точностью до 0,1 нм и практически не зависит от состава пробы, т. е. это устойчивый и аддитивный поисковый признак. Из нескольких сот известных полиаренов в природных и техногенных объектах распространены лишь несколько десятков соединений [291]. Их квазилинейчатые спектры люминесценции хорошо изучены; по атласам [292-294] можно составить небольшую по объему, но достаточно полную базу данных для ИПС. Отработка стратегии поиска на линейчатых спектрах другого типа (например, атомно-эмиссионных, где число линий существенно больше) требует гораздо больших затрат времени.

Идентификация полиаренов в их смесях имеет и важное практическое значение. Многие из них - опасные канцерогены; их обнаружение и определение (вплоть до 10^{-10} %) требуется в анализе воздуха, пищевых продуктов, различных вод, нефтепродуктов и других объектов, где изомеры с различной канцерогенной активностью всегда присутствуют совместно. Полное разделение смесей полиаренов методами ГЖХ или ВЭЖХ, как и специфическое возбуждение люминесценции отдельных компонентов смеси, связано с большими экспериментальными трудностями. Для практики наиболее перспективны методы идентификации полиаренов по неселективно возбуждаемым спектрам неразделенных смесей. На протяжении 80-90-х годов развитие теории идентификации и расширение возможностей компьютерной техники приводило к постоянному обновлению соответствующих БД и ИПС. Примером могут быть ИПС серии “Спектр”, разработанные в ОмГУ для поиска канцерогенных полиаренов (табл.4.4).

В табл.4.4 *предельная сложность* исследуемой смеси охарактеризована максимально допустимым при расшифровке реального спектра пробы числом одновременно

---

[2] Аналогичные, но менее обширные результаты получены при расшифровке линейчатых атомно-эмиссионных спектров. Предварительные результаты были положительны и в случае масс-спектров.

люминесцирующих полиаренов, когда большинство этих полиаренов опознаются правильно. Примером может быть результат расшифровки состава двухкомпонентной смеси (табл. 4.5).

Таблица 4.4.

Основные характеристики ИПС серии "Спектр"

Характеристика	*Спектр--1*	*Спектр-2*	*Спектр-3*	*Спектр-4*
Время разработки	***1980-1982***	***1984-1986***	***1988-1991***	***2000-2001***
Тип ЭВМ	***ЕС-1033***	***ЕС-1045***	***IBM PC/AT***	***Pentium***
Язык программирования	***Fortran IV***	***Fortran IV, PL-1***	***Pascal Turbo***	***C++***
Число спектров в БД	***100***	***150***	***200***	***303***
Время расшифровки, мин:	***<3***	***<2***	***<0,5***	***2-3***
Способ отбраковки:				
по растворителю	-	-	-	+
По "последним линиям"	-	-	+	+
по структуре X_i	-	+	+	+
по способу возбуждения	+	+	+	+
по числу случайных совпадений	-	+	+	–
по критическому уровню сигнала	+	-	-	+
Предельная сложность смеси	***3***	***6***	***10***	***15***

Таблица 4.5

Расшифровка состава бинарной смеси полиаренов с помощью ИПС "Спектр-2" [113]

Компонент	l_i	n_i	$N_{кр}$	S_i	A_i	T_i
Фенантрен	***22***	***14***	***4***	***7,03***	***0,77***	***7,5***
Трифенилен	***26***	***25***	***5***	***3,59***	***0,78***	***4,2***
Хризен	***14***	***4***	***3***	***0,64***	***0,11***	***0,89***

Примечание. Компоненты, у которых $n_i < n_{кр}$, исключены. Названия действительно присутствующих в пробе компонентов (10^{-7} г/мл) выделены курсивом.

Вероятностный критерий, рассчитанный по формуле (4-8), позволил отбраковать 197 из 200 проверенных полиаренов. В списке предположительно идентифицированных остались лишь 2 действительно присутствующих компонента бинарной смеси и один псевдокомпонент, причем сигнал последнего намного ниже, и это может быть учтено пользователем при интерпретации машинного ответа. Однако при анализе более сложных смесей число ложно идентифицированных полиаренов увеличивается. Так, при расшифровке спектра 6-компонентной смеси машинный ответ той же ИПС содержал названия 5 реально присутствующих полиаренов и 7 посторонних соединений, не являющихся их структурными аналогами. Ложные идентификации можно устранять, учитывая дополнительные поисковые признаки, например относительные интенсивности линий или условия возбуждения спектра (см. раздел 4.5). Однако возможен и иной путь - совершенствование способа расчета $n_{кр}$, а именно – учет характера распределения линий в сопоставляемых спектрах. Критическое число совпадений с i-м эталоном вследствие неравномерного размещения линий может оказаться и больше, и меньше, чем при расчете $n_{кр}$ по схеме Бернулли.

Существует эмпирический способ оценки $n_{кр}$, пригодный при любом распределении линий в спектре пробы и легко реализуемый в компьютерном эксперименте [272]. А именно:

1) спектр пробы моделируют набором случайных чисел, получая псевдоспектр;
2) линии псевдоспектра поочередно сопоставляют с линиями всех компонентов, присутствие которых ожидается в пробе, фиксируя совпадения по длине волны в пределах $\pm\Delta\lambda$ и количество таких совпадений по каждому X_i (n_i');
3) моделирование и расшифровку псевдоспектров повторяют k раз;
4) сопоставляют по k однотипным псевдоспектрам полученные значения n_i', находя максимальное значение этой величины (далее n_{ik}');
5) n_{ik}' берут в качестве эмпирической оценки $n_{кр}$ по данному X_i . Оценка будет тем точнее, чем более сходны псевдоспектры со спектром пробы (при условии, что сравнивали

достаточно большое число псевдоспектров);

6) найденное для i-го эталона число совпадений с реальным спектром пробы (n_i) сопоставляют с n_{ik}'. Если, например, в спектре пробы 9 линий оказались совпадающими с спектром антантрена и 2 со спектром пирена, тогда как с псевдоспёктрами той же пробы при том же значении $\Delta\lambda$ антантрен дает самое большее 4, а пирен 3 совпадения, то антантрен следует признать идентифицированным (9>4), а пирен — не идентифицированным (2<3).

Основная проблема при реализации данного способа заключается в получении псевдоспектров, в точности сходных со спектром пробы. Компьютерные программы типа датчика случайных чисел могут обеспечить равномерное, или нормальное, или другое известное распределение линий в псевдоспектрах, но не могут привести к точно такому же эмпирическому распределению, как в спектре пробы. Поэтому в [137] предложен другой способ моделирования спектров — *метод сдвига*, близкий известному математическому приему кросс-корреляции [295]. А именно, ко всем длинам волн линий из спектра пробы добавляют одно и то же произвольное число В, удовлетворяющее условию:

$$|\lambda_2 - \lambda_1| >> |B| > 2\Delta\Delta \qquad (4\text{-}9)$$

Положительные значения В соответствуют небольшому сдвигу всех линий из спектра пробы в сторону больших, а отрицательные — в сторону меньших длин волн. После сдвига спектры содержат столько же линий, что и спектр пробы; имеют точно такое же распределение линий в шкале длин волн, каждая линия относится практически к тому же интервалу длин волн, что и до сдвига, но совпадения со спектрами каких бы то ни было X_i после сдвига становятся чисто случайными. Вероятность совпадения для разных линий компонента X_i может быть совершенно неодинаковой, но в каждом случае она остается практически той же, что до сдвига. Следовательно, статистическое суммирование P_{ij} в ходе компьютерного эксперимента может дать более точную оценку критерия, чем расчет по формулам Бернулли, сделанный в предположении равномерного распределения. Если же линии в сопоставляемых спектрах на самом деле распределены равномерно, то оба способа оценки должны привести к одинаковым значениям критерия. Для разных эталонов значения n_{ik}', полученные при прочих постоянных условиях, оказываются неодинаковыми. Для одного и того же эталона n_{ik}' тем выше, чем больше $\Delta\lambda$ и заселенность спектра пробы, т. е. наблюдается полная аналогия с изменением расчетной величины $n_{кр}$.

Важным условием правильной оценки $n_{кр}$ является обоснованный выбор k - количества генерируемых псевдоспектров. При низких k получали заниженные значения $' n_{ik}$ и поэтому не устраняли ложные идентификации полностью; слишком большие k иногда вели к явно

завышенным оценкам $n_{кр}$, т. е. к отрицательному результату при поиске действительно присутствующих компонентов. Оптимальное значение k можно вычислить, исходя из желательного уровня значимости нуль-гипотезы: $k = \alpha^{-1} - 1$, в частности при $\alpha = 0{,}10$ спектр пробы следует моделировать и расшифровывать 9 раз, при $\alpha = 0{,}05$ - 19 раз [255].

При достаточном быстродействии ЭВМ для точной оценки следует брать очень большое число псевдоспектров (500, 1000), но в качестве критерия тогда надо использовать не максимальное значение n_i', а квантиль полученной совокупности значений n_i', соответствующий желаемому значению $1 - \alpha$. Так, сопоставляя эталонный спектр фенантрена с 100 однотипными псевдоспектрами, компьютер выявил в каждом случае разное число случайных совпадений, но для 95 псевдоспектров это число не превышало 3, и лишь в 5 случаях было больше. Следовательно, 95-процентный квантиль данной выборки равен 3, это число и берется в качестве оценки $n_{кр}$.

Проверка метода сдвига в компьютерных экспериментах показала, что при резко неравномерном распределении линий в спектре пробы метод сдвига и схема Бернулли действительно дают существенно различные оценки $n_{кр}$. В случае спектров НЛ различие иногда достигало 4 -5 единиц. В частности, если расчет критериев по схеме Бернулли вели для широкой области длин волн, а линии в спектрах пробы и эталонов концентрировались в одной и той же узкой области, то расчеты по схеме Бернулли приводили к заниженным значениям $n_{кр}$, а следовательно, и к ложным идентификациям (табл.4.6). Из этой таблицы видно, что кроме действительно присутствующего в пробе 1,12-бензперилена, опознаны еще 2 соединения. Сопоставление же n_i с n_{ik}' (метод сдвига) позволяет получить правильные и однозначные результаты.

Таблица 4.6

Применение разных критериев идентификации
при неравномерном распределении линий в спектре пробы

Эталон	n_i	$n_{кр}$ *при* α = *0,10 (расчет)*	n'_{ik} *(метод сдвига)*
1,12-Бензперилен	*27*	*3 (+)*	*6 (+)*
1-Метилтетрафен	***5***	***1 (+)***	***5***
Флуорантен	***3***	***2 (+)***	***4***
4-Метилтетрафен	***1***	***1***	***2***

Примечание. Проба - 10^{-7} М раствор 1,12-бензперилена, спектр снят по методике [292]. При расшифровке $\Delta\lambda$ = 0,2 нм. Опознанные соединения помечены знаком (+). Данные по всем неопознанным эталонам (кроме 4-метилтетрафена) опущены.

Как видно из табл.4.6, вероятностные критерии подобного типа применимы не только для анализа смесей, но и при опознании чистых веществ, в этом случае обратный поиск становится столь же эффективным, как и прямой. Так, при расшифровке эталонных спектров НЛ с помощью ИПС “Спектр-2” индивидуальные полиарены опознавались всегда правильно и в 90% из 200 проведенных проверок - однозначно, то есть без ложных идентификаций. Примерно такие же результаты были получены при расшифровке спектров, представляющих собой суперпозицию двух-трех эталонных спектров из БД (бинарно кодированных). Однако возможности этого алгоритма пока трудно оценить, так как проверка проведена лишь при использовании небольшой по объему базы данных (200 эталонных спектров). БД по спектрам другого типа (ИК-, масс-, ЯМР и т.п.) содержат гораздо большее число эталонных спектров, поэтому потребуется и большее число последовательных проверок (далее это число обозначается символом Q), что может привести к потере однозначности. Этот вопрос требует дополнительного разъяснения.

А именно, при больших Q существует высокая вероятность (α_Σ) того, что хотя бы по одному из отсутствующих компонентов число совпадений случайно превысит $n_{кр}$, несмотря на низкие значения α по каждому X_i в отдельности.[3] По формуле умножения вероятностей:

$$\alpha_\Sigma = 1 - (1-\alpha)^Q \qquad (4\text{-}10)$$

Из (4-10) следует, что для исключения **всех** ложных идентификаций с заданной

[3] Это положение поясняет простая аналогия: вероятность случайно встретить кого-нибудь из своих знакомых значительно выше вероятности случайной встречи с любым заблаговременно намеченным знакомым. Раззичие вероятностей тем больше, чем больше знакомых!

доверительной вероятностью $P_{иден}$ = 1 - α_{Σ} нужно, чтобы по каждому из проверяемых эталонов значение α не превышало бы 1 - $\sqrt[Q]{\alpha_{\Sigma}}$. При прочих равных условиях значение α должно быть тем меньше, чем больше Q. Для исключения всех ложных идентификаций с 90%-ной доверительной вероятностью величину α в формуле (4-5) надо брать: при Q = 2 на уровне 0,05 (или ниже); при Q = 10 - на уровне 0,01; при Q=50 - на уровне 0,002 и т. д. Критерий $n_{кр}$ будет соответственно повышаться (табл. 4.6), независимо от того, найден ли он чисто расчетным путем или с применением методом сдвига. Аналогичный подход используется и в количественном анализе [296].

Описанный выше способ устранения ложных идентификаций должен быть весьма эффективен при использовании больших БД. Имеется лишь одно серьезное осложнение: при больших Q значения $n_{кр}$ оказываются слишком большими, они начинают превосходить n_i не только для всех отсутствующих (что и требовалось), но и для некоторых из реально присутствующих компонентов пробы. ЭВМ начинает отбраковывать и эти соединения, особенно при неоптимальных условиях возбуждения или регистрации спектра, при наличии сильного диффузного фона в спектре, а также при низкой концентрации компонентов пробы.

Таблица 4.7.

Значения $n_{кр}$, *гарантирующие* ($P_{иден}$ = *0,90) устранение всех случайных ложных идентификаций при разном числе (*Q*) последовательно проверяемых эталонов*

P	Q			
	1	*10*	*100*	*1000*
0,010	***1***	***2***	***3***	***4***
0,100	***6***	***9***	***11***	***12***
0,250	***14***	***17***	***19***	***21***

Естественным способом снижения вероятности ложных идентификаций без повышения $n_{кр}$ в случае большой БД является ее предварительное сокращение, априорная отбраковка всех эталонов, заведомо отсутствующих в пробе. Так, при формировании рабочей библиотеки можно исключить часть эталонных спектров, если:

- все линии эталонного спектра лежат в диапазоне длин волн, не совпадающем со спектром пробы;

- эталонный спектр находится в антистоксовой области по отношению к режиму

возбуждения, использованном при регистрации спектра пробы;

- в структуре эталонного соединения есть элементы, заведомо отсутствующие в пробе, например, галогены, или азот, или сера;

- в спектре пробы не проявилась наиболее интенсивная линия данного эталонного спектра.

Последний фильтр является наиболее эффективным, он позволяет заранее исключать до 80% эталонов. Это предотвращает появление завышенных значений $n_{кр}$, а тем самым пропуск действительно присутствующих компонентов пробы. Кроме того, предварительная отбраковка заведомо отсутствующих соединений позволяет существенно повысить характеристичность признаков остальных предполагаемых компонентов пробы (см. раздел 4.5). Анализ литературы показывает, что предварительная отбраковка может проводиться по многим другим признакам, специфическим для каждого типа спектров. Так, в ИПС "Sampo" при расшифровке гамма-спектров пользователь может учесть до 50 разнородных признаков [297].

4.5. Учет специфичности и относительной интенсивности линий

К сожалению, при расчете вероятностных критериев идентификации по схеме Бернулли пока не удается учитывать характеристичность поисковых признаков, и это вызывает необходимость рассмотрения дополнительных характеристик каждого предполагаемого компонента пробы. Одной из таких характеристик может быть *сигнал присутствия*, другой – *отношение сигнал/шум.* Эти характеристики стоит вычислять лишь для компонентов, прошедших через фильтры предварительной отбраковки и выдержавших сравнение с критерием идентификации, например, удовлетворяющих условию $n_i > n_{kp.}$, а учитываться они должны пользователем лишь при конечной интерпретации поискового (машинного) ответа.

Понятие сигнала присутствия и способы расчета его в обычном и в нормированном виде уже обсуждались в разделе 4.2. Как видно из таблицы 4.5, при расшифровке спектров НЛ величина нормированного сигнала присутствия (A_i), вычисленного по формуле (2) и учитывающего характеристичность отождествленных линий, - для действительно присутствующих компонентов смеси близка к единице, а для псевдокомпонентов стремится к нулю. Отсутствующие в пробе структурные аналоги действительно присутствующих соединений могут давать разные значения A_i, смотря по степени сходства их эталонных спектров.

Оценивая сигнал присутствия, можно учесть не только число возможных межэалонных наложений, но и величину $\mathbf{I_{отн}}$ ***- относительную интенсивность линий***

эталонного спектра, она служит вспомогательной характеристикой. При этом интенсивность линий в спектре пробы не учитывается. Так, в ИПС "Спектр-3" [94] для каждой проверяемой (j-ой) линии i-го эталона характеристичность G_{ij} рассчитывается с учетом числа возможных наложений и объема БД, как описано в разделе 4.2), а затем домножается на $I_{отн}$. Суммирование соответствующих произведений по всем линиям пробы, отнесенным к эталонному спектру X_i , дает обобщенный сигнал присутствия X_i в данной пробе. Суммирование тех же произведений по всем линиям i-го эталона дает максимально возможный уровень сигнала. Нормированный сигнал присутствия (A_i) вычисляется по формуле (4.2). Параметр A_i стремится к 1 при увеличении числа совпадающих линий пробы и эталона, особенно при проявлении в спектре пробы наиболее интенсивных или наиболее специфических линий i-го эталона. Сходный подход применялся и при расшифровке масс-спектров [283]. Учет интенсивности эталонных линий позволяет расширять БД без заметного снижения надежности работы ИПС [94, 298].

Очевидно, для более надежного обнаружения компонентов требуется как можно сильнее повысить их сигнал присутствия. Из формулы (4-1) следует, что этого можно достичь двумя способами. Первый - увеличение S_i за счет роста l_i, то есть введение дополнительных поисковых признаков, - приводит к длительным и трудоемким методикам анализа, громоздкости баз данных, увеличению времени работы ИПС. Так, идентификация вещества по совпадению экспериментальных и табличных индексов удерживания на 10-20 разных хроматографических колонках безусловно надежна, но в реальных условиях такой способ вряд ли применим. Второй способ - повышение специфичности ограниченного набора признаков. Для этого следует устранять возможность межэталонных наложений. На достижение именно этой цели направлены такие известные приемы, как предварительное фракционирование пробы; ввод маскирующих реагентов; изменение условий испытаний (переход к селективным детекторам в ГЖХ, переход в область отпечатков пальцев в ИК-спектрометрии); увеличение точности измерений D_{ij}, и другие приемы. До появления ИПС эти приемы совместно не рассматривались и не имели количественных оценок эффективности.

В качестве примера рассмотрим спектральную идентификации одного из полиаренов - фенантрена - в присутствии других полиаренов. Сопоставление спектров низкотемпературной люминесценции 200 эталонов показывает, что ни одна из линий в эталонном спектре фенантрена не является характеристической, то есть специфической по положению в шкале длин волн (при $\Delta\lambda$= 0,5 нм) в рамках исходной БД. Предварительное

хроматографическое фракционирование пробы - исключение неуглеводородных компонентов, а затем и алкилированных углеводородов - лишь незначительно повышает сигнал присутствия. Спектральное фракционирование за счет изменения условий возбуждения влияет несколько сильнее. Более эффективным приемом является повышение точности измерений - уменьшение $\Delta\lambda$ до 0,1 нм увеличивает S_i втрое. Одновременное же применение всех этих приемов почти устраняет возможность межэталонных совпадений и приводит к повышению сигнала присутствия в 7 раз. Из 19 линий фенантрена 12 становятся специфическими признаками, обеспечивающими его надежную идентификацию не только в чистом растворе, но и при одновременном проявлении в спектре смеси линий 70 других незамещенных полиаренов [13], в том числе ближайших структурных аналогов фенантрена (изомеры, гомологи и т.п.).

Естественно, для эффективной отбраковки заведомо отсутствующих компонентов в базе данных должна быть как можно полнее отражена информация по составу и структуре предполагаемых компонентов, их отнесению к тому или иному классу, условиям регистрации спектра и т.п. Еще лучше было бы использовать для предварительной отбраковки признаки одного вида (например, положение полос в ИК-спектрах), а для окончательной идентификации - другие признаки (например, масс-спектральные) и вероятностные алгоритмы. Однако комбинированные БД пока применяются лишь для опознания и интерпретации спектров чистых веществ, а также для отнесения хроматографических пиков, но не в анализе неразделенных смесей. Это направление исследований представляется очень перспективным.

Количественно оценить достоверность идентификации можно также по отношению сигнал/шум [113]. В качестве оценки шума в ИПС серии “Спектр” берется усредненный сигнал присутствия i-го компонента, рассчитываемый так же, как и S_i, но не по реальному спектру пробы, а по набору псевдоспектров с тем же числом линий, тем же охватываемым спектральным интервалом и приблизительно тем же характером распределения линий внутри этого интервала. Такие псевдоспектры можно генерировать с помощью датчика случайных чисел или по методу сдвига. После сдвига спектра пробы на 5—10 $\Delta\lambda$ в сторону больших или меньших длин волн и повторной его расшифровки сигнал действительно присутствующего компонента должен резко снизиться, и отношение сигнал/шум (T_i) окажется значительно больше единицы. Так как ложные идентификации при расшифровке спектра пробы и псевдоспектров имеют сходную вероятность, то для отсутствующих в пробе компонентов $T_i \approx 1$. Это различие действительно проявляется при расшифровке спектров смесей известного состава (см. табл.4.5).

Расчет параметра T_i по методу сдвига не требует больших затрат машинного времени, поскольку характеристичности эталонных линий не пересчитываются, а поиск совпадений по псевдоспектрам делается лишь для немногих эталонов, оставшихся после предварительного сокращения БД. Отметим, что отношение сигнал/шум для действительно присутствующих компонентов исследуемой пробы существенно зависит от сложности состава пробы и от условий регистрации ее спектра (см. раздел 4.6).

Интересной проблемой, связанной не только с количественной оценкой сигнала присутствия, но и с расчетом критериев, является выбор величины $\Delta\lambda$, то есть полуширины того спектрального окна, в пределах которого регистрируются совпадения линий пробы и эталона. Очевидно, что эта величина должна задаваться пользователем с учетом воспроизводимости измерения длин волн (на данном приборе и при использовании данного режима регистрации спектра), должна быть учтена и точность табличных данных. Не исключено, что для разных диапазонов длин волн и даже для разных эталонов целесообразно брать различные значения $\Delta\lambda$, оптимизированные таким образом, чтобы уменьшить суммарную вероятность идентификационных ошибок 1-го и 2-го рода. Для хроматографического анализа такой подход оказался весьма плодотворным [254]. Однако в спектральном анализе расчеты критериев и сигналов присутствия при переменной величине $\Delta\lambda$ неизбежно приведут к крайне сложным алгоритмам расчетов, такой подход вряд ли осуществим на практике.

С другой стороны, вызывает сомнения сама идея подсчета спектральных совпадений. Обычно две линии считаются либо совпадающими, либо не совпадающими. Это слишком жесткая логика [79]. Во многих работах делались попытки использовать при оценке близости спектров другие подходы, например, использовать логику нечетких множеств [4, 299]. Применительно к спектральному анализу смесей по линейчатым спектрам это означает отказ от подсчета спектральных совпадений и оценки критического числа совпадений, так как такие подсчеты не учитывают небольших различий в положении линий пробы и эталона (в пределах $\pm\ \Delta\lambda$), не учитываются и линии пробы, близкие к линии эталона, но все же находящиеся за пределами вышеуказанного окна. В качестве примера можно привести алгоритм ИПС “Арфа”, который не связан с подсчетом совпадений. Спектральные кривые пробы и проверяемого эталона нормализуются, совмещаются, а затем в качестве сигнала присутствия отыскивается доля площади, общая для обоих спектров [300]. Но этот алгоритм позволяет анализировать лишь сравнительно простые смеси (2-3 компонента с концентрациями одного порядка).

Известно, что погрешности в оценке положения спектральных линий в большинстве случаев имеют нормальное распределение. Это позволяет рассчитать сигнал присутствия

любого X_i в спектре смеси с помощью функций Лапласа (идея выдвинута В.А.Топчием в 2000 г.). Степень близости i-го эталонного спектра и спектра пробы можно оценить по величине функционала

$$F_i = \sum_j \left(1 - 2\Phi \frac{\left|\lambda_{ij} - \mu_j\right|}{\sigma} \right) \qquad (4\text{-}11),$$

где λ_{ij} и μ_j - длины волн сопоставляемых линий из спектра эталона и пробы, а σ - стандартное отклонение при измерении длин волн. Для простоты можно ограничиться попарным сопоставлением ближайших линий. Суммирование функций Лапласа ведется по всем линиям эталонного спектра. Тогда при идеальном совпадении спектров функционал равен числу линий эталонного спектра, а по мере нарастания отличий в сопоставляемых спектрах значение функционала падает до нуля, но не дискретно, как значение сигнала присутствия, вычисленное по формуле (4-1), а непрерывно. Специфичность и относительная интенсивность соответствующих линий эталонного спектра могут быть весовыми показателями при суммировании слагаемых по формуле (4-11), а отношение полученной суммы к ее максимально возможному значению даст нормированный сигнал присутствия. Критическое же значение сигнала присутствия можно получить при многократном генерировании псевдоспектров (например, по методу сдвига), их расшифровке и подсчете сигналов присутствия для каждого псевдоспектра. Эта идеология заложена в качестве основы новой ИПС “Спектр-4”.

В заключение этого раздела стоит перечислить еще несколько интересных и пока нерешенных проблем теоретического характера, связанных с методологией обратного поиска. Как добиться однозначного опознания компонента смеси в условиях неслучайных совпадений, то есть отличить присутствующие вещества от их структурных аналогов? Как оценить вероятность случайного недооткрытия компонента? Как использовать неустойчивые поисковые признаки, например относительные интенсивности спектральных линий? Последнее важно не только для расшифровки линейчатых спектров: поисковые признаки вообще не являются абсолютно независимыми, устойчивыми и аддитивными, и именно это обстоятельство вынуждает вводить операции предварительного разделения компонентов.

4.6. Анализ модельных и реальных смесей полиаренов с помощью ИПС “Спектр”

В качестве примера рассмотрим аналитические возможности ИПС “Спектр-3”. Проверка работы этой ИПС проводилась несколькими способами: в компьютерных экспериментах, на модельных однокомпонентных и многокомпонентных растворов разной концентрации и, наконец, при расшифровке состава различных природных и техногенных объектов. В последнем случае результаты сопоставлялись с данными, полученными с помощью других методов (например, ВЭЖХ). На каждом из этапов расшифровывали 50-100 однотипных спектров. Кроме того, с целью оптимизации сопоставляли результаты расшифровки одних и тех же спектров в разных режимах (варьирование $\Delta\lambda$, $\alpha_{кр}$ и т.п.). Всего в ходе проверки было расшифровано свыше 800 спектров.

При использовании оптимальных режимов результаты сводятся к следующему.

1). При вводе на расшифровку случайных наборов чисел, внешне напоминающих спектры НЛ (“**спектры белого шума**”), никакие полиарены не опознаются.

2). При вводе **эталонных спектров полиаренов, включенных в БД**, все соединения опознаются правильно, причем в 87% случаев однозначно, то есть машинный ответ содержит только одно название. Немногочисленные псевдокомпоненты оказываются структурными аналогами опознаваемых соединений, например, имеют то же ароматическое ядро, но другой заместитель; или тот же заместитель, но в другом положении.

3). При вводе **спектров индивидуальных полиаренов, отсутствующих в БД**, лишь в 6 случаях из 50 опознаны структурные аналоги, например, 2-метилпирен вместо 1-этилпирена. В остальных случаях никакие соединения не опознаются. Этот результат очень важен, так как ИПС данной серии предназначены не для исследования строения вновь синтезированных веществ, а для поиска и безошибочного опознания индивидуальных канцерогенов. Отметим, что структура многих канцерогенных полиаренов весьма близка к структуре других, более распространенных, но безвредных соединений. Идентификация вместо канцерогенов их структурных аналогов недопустима (как и противоположная ошибка!). Впрочем, однозначность отбора поддается регулировке - при увеличении $\Delta\lambda$ и $\alpha_{кр}$, а также вследствие других изменений режима работы ИПС частота появления структурных аналогов (отсутствующих в пробе) становится гораздо больше.

4). Следующий этап проверки - **ввод реальных спектров НЛ индивидуальных полиаренов**. Соответствующие соединения представлены в БД, но для расшифровки в ЭВМ вводятся не эталонные спектры, а спектры тех же соединений, заново снятые на различной аппаратуре при разных концентрациях (в интервале 10^{-5} - 10^{-9} г/мл) или взятые из литературных источников. Несмотря на отличия расшифровываемых спектров от эталонных, практически всегда соответствующий полиарен опознавался правильно и в 70-80% случаев однозначно. Его сигнал присутствия был заметно меньше, чем при опознании

того же вещества по эталонному спектру (то есть A_i < 1), это указывает на уменьшение достоверности идентификации. Степень снижения достоверности зависит в основном от техники регистрации реального спектра, а также от числа линий в эталонном спектре и от задаваемого при расшифровке параметра $\Delta\lambda$. При превышении некоторого предела $(\Delta\lambda > \Delta\lambda_{крит})$ индивидуальные полиарены вообще перестают опознаваться. Как и следовало ожидать, наилучшие результаты получаются, если параметр $\Delta\lambda$ совпадает с реальной погрешностью измерения длин волн, а условия регистрации расшифровываемого и эталонного спектра совпадают [282]. По мере снижения концентрации в спектре раствора наблюдается все меньшее число линий опознаваемого соединения, поэтому сигнал присутствия падает ($A_i \rightarrow 0$).

5). Традиционным способом проверки ИПС, предназначенных для качественного анализа смесей, является конструирование **суперпозиций из нескольких случайно отобранных из БД эталонных спектров** и их последующая расшифровка [299-300]. В модельном спектре присутствуют все линии каждого из Y использованных эталонов. Таким образом моделируются реальные спектры аддитивных Y-компонентных смесей с близкими концентрациями компонентов. В каждом случае ИПС опознает y_1 соединений правильно и y_2 соединений ошибочно. Конструирование и расшифровку повторяли по 10 раз для каждого значения Y. На рис.4.2 приведены усредненные результаты этих экспериментов.

Видно, как при одинаковом режиме расшифровки увеличение сложности смеси (рост Y) постепенно ухудшает правильность опознания (y_1 < Y) и усиливает неоднозначность результатов (рост y_2). Эти эффекты нельзя объяснить погрешностями регистрации спектра или неаддитивностью, ухудшение результатов идентификации при переходе к более сложным смесям может быть связано только с повышением заселенности спектра пробы. Соответственно критическое число случайных совпадений увеличивается до такой степени, что некоторые компоненты перестают опознаваться. Действительно, детальное исследование показывает, что

обычно не опознаются соединения, имеющие всего 3-5 линий в эталонном спектре. Ложно идентифицированными чаще оказываются структурные аналоги “присутствующих” полиаренов, но среди псевдокомпонентов были и соединения с несходной структурой.

Полученные результаты в целом удовлетворительны: даже в случае 15-компонентных смесей правильно опознаются 13-14 полиаренов, то есть почти все “присутствующие”, а также 4-5 их структурных аналогов. Такие результаты получали, если требовали 95%-ой надежности идентификации и не учитывали число проверяемых эталонов. Переход к 99-процентной надежности и применение формулы (10) существенно уменьшает y_2 (иногда до 0), но одновременно на 20-30% снижается и число верно опознанных соединений.

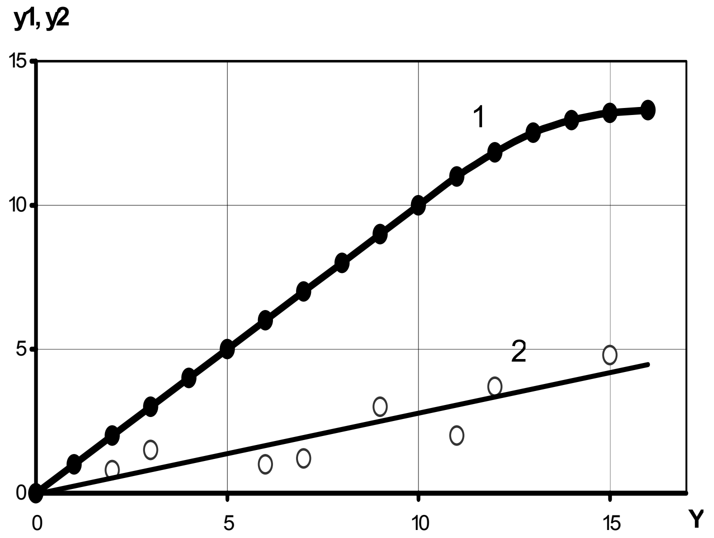
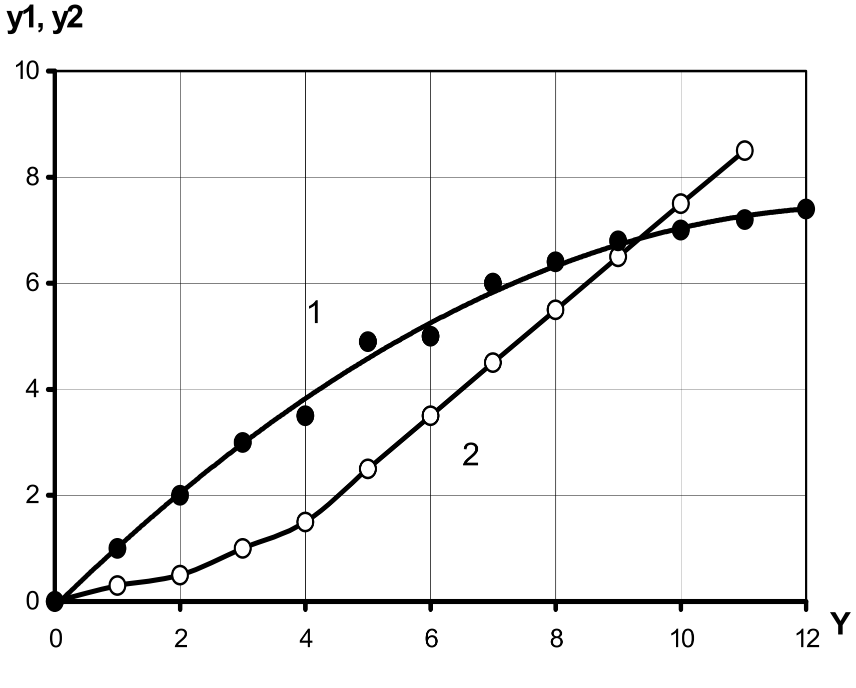


Рис.4.2. **Результаты проверки ИПС “Спектр-3” при расшифровке модельных спектров для аддитивных смесей различной сложности.**

Примечание. Обозначения см. в тексте.

6). **Спектры реальных смесей,** содержащих Y индивидуальных полиаренов в близких концентрациях (10^{-6} - 10^{-7} г/мл), снимали в условиях неселективного возбуждения. Затем проводили расшифровку спектров и определяли y_1 и y_2, как в предыдущей серии.



Результаты, естественно, оказались менее впечатляющими, чем при расшифровке модельных спектров (рис.4.3). Сопоставление результатов расшифровки модельных и реальных спектров смесей показывает, что неаддитивность реальных спектров НЛ и погрешности их регистрации приводят не столько к повышению y_2, сколько к снижению y_1, некоторые из действительно присутствующих соединений не опознаются. Для остальных существенно снижается параметр A_i и отношение сигнал/шум. Результаты можно считать удовлетворительными вплоть до Y = 9-10, когда число правильно опознанных соединений становится равным числу ложно идентифицированных. Для сравнительно простых смесей неудовлетворительные результаты наблюдались, когда соотношение концентраций компонентов было неблагоприятным (например, не опознавался компонент, концентрация которого была на 2 порядка меньше концентрации других компонентов той же смеси). При компьютерном качественном анализе смесей переменного состава аналогичные результаты получены и с помощью ИПС “Арфа” [300].

Рис.4.3. **Идентификация полиаренов по реальным спектрам смесей разной сложности**

Режим расшифровки и обозначения соответствуют рис.4.2.

ИПС “Спектр-3” успешно применяли и при расшифровке спектров смесей с числом компонентов, большим 10, или при опознании микрокомпонентов. Средством для решения таких задач является спектральное фракционирование, то есть спектры пробы надо снимать несколько раз при разных $\lambda_{возб}$ и проводить расшифровку каждого из полученных спектров по отдельности. Присутствующим следует считать соединение, которое с заданной надежностью было опознано хотя бы одному из этих спектров. Примером может быть анализ 12-компонентной модельной смеси (табл.4.8). Используя всего 4 режима возбуждения, мы верно опознали все присутствующие соединения, а ложно опознано было всего лишь два полиарена [13]. Это объясняется тем, что одновременно люминесцировало не более 8 компонентов смеси.

Таблица 4.8

Идентификация индивидуальных полиаренов в 12-компонентной модельной смеси

в условиях спектрального фракционирования

возб, нм	Действительные компоненты смеси												Ложно опознанные полиарены
		н	А		л	П	еП	л	БА	Пл	Бп		
90							-		+			-	**3-МБА**
37							-		+			-	**6-МБА**
70							-		-			-	**Нет**
05							-		-			-	**Нет**
							-		+			-	++

Примечание*: условно обозначены названия полиаренов*: Н - нафталин, Фн - фенантрен; БА - 1,2-бензантрацен; П - пирен; Фл - флуорантен; БП - 3,4-бензпирен; БеП - 1,2-бензпирен; Пл - перилен; ДБА - 1,2-3,4-дибензантрацен; БПл - 1,12-бензперилен; ДБп - 1,2-3,4-дибензпирен; К - коронен, 3-МБА - 3-метилбензантрацен; 6-МБА - 6-метилбензантрацен.

7). Заключительный этап проверки ИПС - **анализ многокомпонентных смесей априорно неизвестного состава**, имеющих природное или техногенное происхождение. Объекты анализа - сланцевая смола, технический углерод, подземные воды, сточные воды нефтеперерабатывающих предприятий, автомобильный выхлоп и т.п. Анализ тех же проб с помощью других методов проводлился параллельно и независимо в лабораториях других организаций, в том числе зарубежных. При этом применяли многоступенчатое хроматографическое разделение смесей вплоть до индивидуальных соединений, проверяли характеристики удерживания, а затем исследовали спектры узких фракций, например, спектры поглощения в УФ-области или спектры НЛ, полученные при селективном возбуждении. Результаты качественного анализа в основном совпадают [15, 94 и др.]. Примером могут быть данные по составу автомобильного выхлопа. Присутствие пирена, метилпиренов, антантрена и 1,12-бензперилена в пробе было надежно установлено шведскими исследователями методом ВЭЖХ и по виду спектра НЛ соответствующей хроматографической фракции (методика описана в работе [301]). Те же соединения были опознаны с помощью ИПС серии “Спектр” в ОмГУ [137], различие наблюдалось только в одном случае из шести (компьютер установил присутствие метилпирена, но указал на другое положение заместителя).

Несколько большие расхождения выявлены при параллельном исследовании сточных вод нефтеперерабатывающих предприятий [302]. В условиях спектрального фракционирования в неразделенном экстракте одной из проб было опознано с помощью ИПС серии “Спектр” 39 соединений, в том числе алкилированных (время анализа - 3 часа).

С помощью гораздо более длительной методики, предусматривающей хроматографическое фракционирование экстракта и поиск характерных пиков в спектрах НЛ, было найдено 15 соединений из того же списка (проверка присутствия алкилированных соединений в стадартной методике не предусмотрена). Менее чувствительный метод ВЭЖХ со спектрофотометрическим детектированием позволил опознать в той же пробе (по индексам удерживания и УФ-спектрам) лишь 9 полиаренов.

Положительные результаты проверки ИПС "Спектр-3" по вышеописанной схеме позволили использовать ее на кафедре аналитической химии Омского госуниверситета для серийных анализов различных объектов. В настоящее время в ОмГУ по той же схеме идет многостадийная проверка новой системы "Спектр-4", рассчитанной на применение более современной вычислительной техники и гораздо более полной базы данных. Установлено, в частности, что с помощью новой ИПС при расшифровке модельных спектров НЛ аддитивных смесей можно устанавливать состав проб, содержащих 15-20 одновременно люминесцирующих компонентов, то есть анализировать существенно более сложные объекты, чем с помощью ИПС "Спектр-3".

В заключение рассмотрим некоторые ИПС, реализующие ту же методологию обратного поиска, те же алгоритмы расчета сигнала присутствия и вероятностных критериев идентификации, но предназначенные для расшифровки линейчатых спектров другого типа, например, атомно-эмиссионных. Естественно, переход к новому типу спектров в каждом случае требует создания специальной базы данных, но в самой программе расшифровки спектра пробы приходится вводить сравнительно небольшие изменения. Наибольший интерес представляет ИПС "Аргус-2" (ОмГУ, 1993), реализованная на компьютерах типа IBM PC/AT [303]. Время расшифровки одного спектра не превышает минуты.

Как показано в разделе 4.4, возможность применения вероятностных алгоритмов в расшифровке линейчатых спектров смесей определяется заселенностью спектров эталонов и пробы. При очень большом числе линий в спектре пробы (N) или эталона (l_i), а также при неточном измерении положения линий заселенность сопоставляемых спектров и вероятность случайных совпадений стремятся к единице, и требуемая надежность идентификации не достигается. В базе данных, с которой работает ИПС "Спектр-3", эталонные спектры НЛ в интервале 350-650 нм содержат до 30 линий (в среднем 9,5 линий на один спектр). С учетом точности измерений максимальная заселенность эталонных спектров НЛ составляет величину порядка 2 % (на деле несколько больше из-за неравномерного размещения линий). Довольно низкая заселенность эталонных спектров

НЛ и обеспечивает возможность применения вероятностных алгоритмов даже при расшифровке спектра, содержащего все линии 10-20 индивидуальных полиаренов.

Теперь рассмотрим, можно ли рассчитывать на эффективное применение вероятностных алгоритмов при расшифровке атомно-эмиссионных спектров. При создании ИПС серии “Аргус” в качестве информационной основы был использован известный справочник Зайделя [304]. В БД были включены все линии 70 элементов (для дугового возбуждения) в интервале 250-350 нм, за исключением очень слабых (для которых $I_{отн.}$ < 0,001 I_{max}). В среднем в БД содержится 82 линии на 1 эталонный спектр элемента. Максимальное число линий одного элемента не превышает 500. Для каждой линии указано положение с точностью до 0,001 нм. Таким образом, заселенность эталонного спектра любого элемента не превышает 1%. Так как его линии размещены более равномерно, чем линии в эталонных спектрах НЛ, вероятность случайных совпадений снижается еще сильнее. Следовательно, вероятностные алгоритмы вполне применимы и в атомно-эмиссионном спектральном анализе. Более того, можно ожидать, что с помощью таких алгоритмов можно будет анализировать более сложные смеси, чем рассмотренные выше смеси полиаренов.

При работе с ИПС “Аргус” в компьютер должны быть введены спектр пробы (как последовательность чисел, соответствующих длинам волн всех линий), погрешность измерения длин волн, а также названия заведомо присутствующих и заведомо отсутствующих элементов. Алгоритм работы ИПС включает предварительную отбраковку элементов, для которых в спектре пробы не найдены три самые интенсивные линии эталонного спектра (эти “последние линии” выделены в отдельный файл). Новым приемом по сравнению с ИПС “Спектр-3” является возможность предварительного удаления из спектра пробы всех линий, которые можно отнести к заведомо присутствующим элементам (например, линий железа при расшифровке состава сталей). При этом теряется и часть линий других элементов, но это не мешает их опознанию, так как линии, совпадающие с какой-нибудь из последних линий, не удаляются. Другие отличия ИПС серии “Аргус” от описанных выше ИПС серии “Спектр” представляются второстепенными.

Проверка ИПС “Аргус-2” проводилась в основном по вышеприведенной схеме. Никакие элементы не опознавались, если на расшифровку вводили спектр белого шума или эталонный спектр элемента, отсутствующий в БД. ИПС правильно и однозначно опознавала любой элемент по эталонному спектру из БД или по модельному спектру аддитивной смеси. Для всех отсутствующих в пробе элементов число совпадений оказывалось меньше, чем критическое число совпадений, рассчитанное по формуле (4-8), и они отбраковывались даже при P = 0,90. Применению вероятностных алгоритмов в атомно-эмиссионном

спектральном анализе благоприятствует то, что в данном случае, в отличие от спектроскопии НЛ, идентификации действительных компонентов пробы не мешает опознание их структурных аналогов, то есть ложные идентификации могут иметь только случайный характер.

Поскольку другие линейчатые спектры (например, рентгенофлуоресцентные, спектры РФА и т.п.) однотипны с атомно-эмиссионными спектрами элементов и отличаются еще меньшей заселенностью, то нет сомнений, что методология обратного поиска и вероятностные алгоритмы могут быть применены при расшифровке и этих спектров.

ЗАКЛЮЧЕНИЕ

Приведенные в этой книге примеры качественного анализа реальных объектов наглядно показывают, что лимитирующей стадией, своеобразным “слабым звеном”, пока что является интерпретация экспериментальных данных. Наиболее яркий пример – хроматография с масс-спектрометрическим детектированием. Здесь в ходе анализа смеси неизвестного состава, содержащей десятки индивидуальных соединений, удается (иногда всего за несколько минут) охарактеризовать ее компоненты не только временами удерживания, но и чрезвычайно информативными масс-спектрами. Последующая интерпретация такого объема информации без применения современных компьютерных технологий потребовала бы недель или даже месяцев напряженной работы высококвалифицированных специалистов. Но и компьютерные технологии сегодня не гарантируют безошибочной и однозначной идентификации всех компонентов пробы. Принятие окончательного решения обычно предоставляется пользователю, причем предполагается наличие должной квалификации, опыта и развитой интуиции. Все это побуждает стремиться к созданию все более совершенных автоматизированных средств интерпретации экспериментальных данных, ориентированных на массового пользователя-аналитика и позволяющих быстро получать правильные и однозначные результаты. В этой книге рассмотрена лишь часть из этих средств, а именно - фактографические базы данных с соответствующим информационно-поисковым и информационно-логическим программным обеспечением.

Альтернативный подход - метод искусственного интеллекта, реализуемый на ЭВМ соответствующими экспертными системами. Например, системы DENDRAL [158], CHEMICS [159], PACTP [160], X-PERT [161]) моделируют присущий исследователю-спектроскописту путь решения данной задачи. Информационную основу экспертных систем составляют базы знаний, сформированные путем анализа научной литературы и отражающие многолетний опыт исследований в соответствующей области. Таким образом, фактографические данные в этом методе также используются, но не прямо, а опосредованно, в виде заранее найденных спектроструктурных корреляций. Эффективность решения задач во многом определяется "качеством" баз знаний, и в этом отношении преимущество может быть отдано системам, содержащим знания, специфические для определенных классов органических соединений. Экспертные системы помогают исследователю установить полную структуру органического соединения, в том числе впервые синтезированного, выявить ранее неизвестные спектроструктурные корреляции и решить ряд других уникальных по своей сложности задач теоретического характера [4]. Однако этот метод не ориентирован на практику массового анализа, в частности, на быстрое и надежное обнаружение известных компонентов по спектру пробы; он не позволяет интерпретировать хроматографические данные, а также анализировать неразделенные смеси. Поэтому для практических целей методы компьютерного качественного анализа, основанные на применении информационно-поисковых систем и описанные в настоящей

книге, имеют значительные преимущества, обусловленные более мощной информационной поддержкой. С помощью БД и ИПС можно решать не только задачи информационно-справочного и чисто аналитического характера, но и изучать особенности спектрального поведения отдельных классов органических соединений, объективно формировать статистически обоснованные спектроструктурные корреляции и т.п. Естественно, этот метод имеет не только достоинства, но и недостатки, и ограничения; они отражены и в этой книге. Нашей целью было объективное изложение возможностей компьютеров в качественном анализе, поскольку для достижения успеха в равной степени важны и хороший аналитик, и оказывающий ему помощь компьютер с соответствующим банком данных и программным обеспечением.

Анализ монографий, обзоров и многочисленных оригинальных публикаций по компьютерной идентификации веществ с применением БД и ИПС, а также собственный опыт авторов позволяют придти к следующим выводам:

- надежность компьютерной идентификации вещества во всех случаях зависит от полноты и достоверности имеющейся в БД информации по поисковым признакам разных органических соединений. Вот почему важнейшей заботой всего сообщества аналитиков (спектроскопистов, хроматографистов и т.д.) должно стать создание и пополнение соответствующих баз данных;
- для повышения надежности обнаружения желательно одновременно использовать несколько разнотипных БД, например по масс-спектрам и хроматографическим индексам удерживания;
- при прочих равных условиях успех компьютерной идентификации зависит от числа поисковых признаков в эталонном спектре опознаваемого вещества (этим фактором определяется, в частности, большая надежность масс-спектрометрической идентификации по сравнению с опознанием веществ по характеристикам удерживания), от воспроизводимости и селективности этих признаков, от состава пробы (чистое вещество или смесь), а также от выбора оптимальных алгоритмов и критериев идентификации;
- во всех случаях необходимо безошибочно регистрировать в стандартных условиях спектр пробы (хроматограмму и т.п.) и получить всю возможную априорную информацию о пробе. Это позволяет вести предварительную отбраковку веществ, заведомо отсутствующих в пробе, и формировать некоторую выборку "наиболее подозреваемых" эталонов, что не только сокращает затраты времени на сопоставление спектров, но и повышает надежность конечных результатов анализа;
- дальнейший ход компьютерного качественного анализа с применением БД и ИПС зависит, прежде всего, от наличия в БД эталонного спектра опознаваемого вещества. Специфика поисковых признаков (спектральных или хроматографических), состав пробы (чистое вещество или смесь) и другие факторы имеют в этом отношении гораздо меньшую значимость;
- компьютерный качественный анализ требует осознанного выбора критериев при принятии решений. Наиболее перспективным и теоретически обоснованным подходом представляется использование критериев, рассчитанных методами теории вероятностей. В частности, целесообразно применение алгоритмов, основанных на постулируемых вероятностях двух возможных ошибок (ложной идентификации отсутствующего вещества и пропуска присутствующего);
- при наличии в БД эталонного спектра опознаваемого вещества должно быть проведено последовательное сопоставление признаков пробы и эталонных спектров по всем возможным компонентам данной пробы. Результатом должна быть количественная оценка нормированного сигнала присутствия (степени совпадения спектров) с последующим применением подходящих критериев идентификации;

- если в БД эталонный спектр опознаваемого вещества отсутствует, то его предполагаемое строение может быть определено в результате сопоставления структурных формул всех найденных в БД спектральных аналогов. При этом, как и в случае анализа многокомпонентных смесей, используются гораздо более сложные алгоритмы, чем обычные алгоритмы информационного ("библиотечного") поиска. Они связаны с такими разделами математики, как теория графов, математическая логика, теория вероятностей, теория информации и др. Поэтому не вполне правильно называть соответствующие пакеты компьютерных программ *информационно-поисковыми* системами; более точными являются термины "информационно-аналитическая", "интерпретирующая" и даже "интеллек-туальная система".

Следует подчеркнуть, что в области качественного анализа необходимо создание единой и официально закрепленной системы терминов. Сегодня, к сожалению, разные авторы зачастую используют разные термины для обозначения одного и того же понятия - и, что еще хуже, одним словом характеризуют самые разные понятия. Понимание терминов не должно зависеть от вида качественного анализа (элементный, молекулярный и т.п.), от используемого метода (МС, ГЖХ и т.п.), от степени применения ЭВМ, а также от того, идет ли речь об опознании индивидуального вещества или о качественном анализе объекта неизвестного состава [10,13]. По-видимому, система терминов, предложенная в настоящей книге, могла бы стать основой для обсуждения этой проблемы заинтересованными специалистами (например, в рамках Научного Совета по аналитической химии РАН или в его терминологической комиссии). Очевидно также, что методы качественного анализа должны рассматриваться с единых методологических позиций и иметь специфический математический аппарат. Нам представляется, что основой для его разработки должны стать теория вероятностей (см. гл.3-4 настоящей книги), математическая логика и теория информации. Добиваться надежной идентификации своих объектов приходится не только химикам-аналитикам, но и криминалистам, микробиологам, специалистам по аэрофотосъемке, технологам, искусствоведам и многим другим специалистам, которые также заинтересованы в создании общей теории и разработки соответствующего математического аппарата.
Безусловно, и развитие информационно-поисковых систем, и применение ИПС в аналитических лабораториях находятся пока что в начальной стадии, несмотря на то, что основные алгоритмы поиска информации в базах данных апробированы и отработаны еще в 80-ые годы. Сегодня активно идет наращивание объемов уже существующих баз данных и расширение круга пользователей соответствующих ИПС. В этой связи хотелось бы обратить внимание на следующее обстоятельство. Увеличение объема БД означает, что в их состав включается все больше записей о соединениях, имеющих (случайно или по объективным причинам) весьма сходные эталонные спектры. В результате характеристичность индивидуальных поисковых признаков постепенно падает, растет вероятность ложных идентификаций. Конечно, наращивание объемов БД и поддержание их в актуализированном состоянии чрезвычайно дороги, и мы вряд ли в ближайшее время столкнемся с опасностью "вырождения поисковых признаков". Тем не менее, стоит заранее указать несколько направлений будущих исследований, нацеленных на предотвращение этой опасности:

- дальнейшее совершенствование инструментальных методов, стандартизация методик регистрации эталонных спектров (хроматограмм и т.п.), борьба за их качество, широкое участие аналитического сообщества в накоплении, систематизации и проверке данных, межгосударственная кооперация в области представления данных и т.п.;
- разработка надежных методов прогноза и точного теоретического расчета аналитических характеристик, разработка эффективных средств представления и манипулирования структурными данными в информационно-логических системах; развитие методов качественного анализа в отстутствие эталонов;

- совершенствование ИПС, способных учитывать характеристичность поисковых признаков и оценивать достоверность результатов собственной работы на этой основе.

Можно с уверенностью утверждать, что в ближайшие годы информационные технологии будут все шире проникать в практику качественного анализа, охватывая все более широкий круг аналитических задач и методов. Авторы надеются, что вместе с читателями книги станут свидетелями новых впечатляющих результатов в этой быстро развивающейся области.

Литература

1 ***Основы аналитической химии (в двух книгах). Под редакцией Ю.А.Золотова. М.: Высшая школа. 1999. Т.2. С.431.***

2 **Ляликов Ю.С., Клячко Ю.А.** ***Теоретические основы современного качественного анализа. М.: Химия. 1978. 312 с.***

3 *Шрайнер Р., Фьюзон П., Кертин Д. и др*. Идентификация органических соединений. М.: Мир. 1983. 704 с.

4 *Эляшберг М.Е., Грибов Л.А., Серов В.В.* Молекулярный спектральный анализ и ЭВМ. М.: Наука. 1980. 307 с.

5 *Мильман Б.Л., Конопелько Л.А.* // Зав. лаборатория. 1999. Т.65, № 1. С.58-59.

6 *McLafferty F.W.* Interpretation of mass-spectra. Reading (USA): Addison-Wesley. 1973. 98 p.

7 *Gray N.A.*. Computer-Assisted Structure Elucidation. New-York: Wiley & Sons. 1986, 536 p.

8 Computer-Supported Spectroscopic Databases / Ed. J.Zupan. Chichester: E.Horwood. 1986, 344 p.

9 Computing Application in Molecular Spectroscopy / Ed. W. George, D. Stieele. Cambrige, U.K.: Royal Soc. of Chem. 1995, 236 p.

10 *Вершинин В.И.* // Журн. аналит. химии. 2000. Т.55, № 5. С.468-476.

11 *Вигдергауз М.С., Семенченко Л.В., Езрец В.А. и др.* Качественный газохроматографический анализ. М.: Наука, 1978. 244 с.

12 *Попов А.А.* // Химическая энциклопедия. М.: Сов. энциклопедия. 1990. Т.2. С.346.

13 *Вершинин В.И.* // Химия в интересах устойчивого развития. 1995. Т.3, N.3. С.245-252.

14 *Дерендяев Б.Г., Лебедев К.С.* / В кн. Математические методы и ЭВМ в аналитической химии. Проблемы аналитической химии. Т.9. М.: Наука, 1989. С.110-123.

15 *Потапов В.М., Розенман М.И., Кочетова Э.К., Покровский Б.И.* Поиск химической информации. М.: МГУ. 1990. 174 с.

16 *Хуторецкий В.М.* Общие представления о поиске научно-технической информации в режиме онлайн. Базы данных STN International в теледоступе. М.: МИЦ РАН–STN при

ИОХ РАН. 2000, 40 с.
17 Registry of Mass Spectral Data./ Ed. by E.Stenhagen, S.Abramson, F.W.McLafferty. New York a.o.: John Wiley. 1974. V.1-4.
18 EPA/NIH Mass Spectral Data Base / Ed. by S.R.Heller, G.W.A.Milne. Washington D.C.: National Bureau of Standards. 1978. V.1-4.
19 EPA/NIH Mass Spectral Data Base. Supplement 1,2 / Ed. by S.R.Heller, G.W.A.Milne. Washington D.C.: National Bureau of Standards. 1980 (Suppl.1), 1983 (Suppl.2).
20 ***McLafferty F.W., Stauffer D.B.*** **The Wiley / NBS Registry of Mass Spectral Data / New York a.o.: Wiley. 1989. V. 1-7.**
21 Eight Peak Index of Mass Spectra: Essential Data from 66720 Mass Spectra.-3rd ed./ Nottingham: The Royal Soc. of Chem. 1983. V.1-3.
22 Mass Spectra of Organic Compounds / Ed. by B.H. Kenett. Melbourne. 1977-1982, V. 1-10.
23 CRC Handbook of Mass Spectra of Drugs / Ed. by I. Sunshine a.o. Boca Raton: CRC Press. 1984, 457 p.
24 A Compilation of Mass Spectra of Drugs / Ed. by R.E. Ardrey a.o. London: Pharmaceutical Press. 1985.
25 Атлас масс-спектров органических соединений / Под ред. В.А.Коптюга. Новосибирск: НИОХ СО РАН. 1977-1988, вып.1-21.
26 Каталог сокращенных масс-спектров / Под ред. А.М. Колчина, Новосибирск: Наука. 1981. 188 с.
27 *Millard B.J.* in Application of Mass Spectrometry to Trace Analysis / Ed. by S.F. Facchetti. Amsterdam. 1982. P. 163.
28 *Полякова А.А., Хмельницкий Р.А.* Масс-спектрометрия в органической химии. Л.: Химия. 1972. 368 с.
29 *Хмельницкий Р.А., Бродский Е.С.* Хромато-масс-спектрометрия. М.: Химия. 1984. 216 с.
30 *Зенкевич И.Г., Иоффе Б.В.* Интерпретация масс-спектров органических соединений. Л.: Химия. 1986. 175 с.
31 *McLafferty F.W., Stauffer D.B.*, Twiss-Brooks A.B., Loh S.Y. // J. Am. Soc. Mass Spectrom., 1991. V.2, №5. P. 432-437.
32 *Davies A.N., McIntyre P.S.* in Computer Application in Molecular Spectroscopy / Ed. by W.O. George and S. Steele. Cambrige, U.K.: Royal Soc. of Chem. 1995. P. 41-59.
33 *Dillard J.G., Heller S.R., McLafferty F.W., Milne G.W.A., Venkataraghavan R.* // Org. Mass Spectrom., 1981. V.16, №1. P.48-49.
34 *Derendjaev B.G.* // European Spectr. News. 1985, № 59. P.17-18.
35 *Saeki S., Yamamoto O.* // Codata Bull.. 1981. V.40, P.53-56.

36 *Беллами Д.* Инфракрасные спектры сложных молекул. М.: зд. иностран. лит., 1963. 590 с.
37 *Guelachvili G., Rao K.N.* Handbook of Infrared Standards. London: Academic Press Inc., 1986. 852 p.
38 *Dolphin D., Wich A.* Tabulation of Infrared Spectral Data. New-York:Wiley & Sons, 1977. 549 p.
39 *Socrates G.* Infrared Characteristic Group Frequencies. Tables and Charts. New-York: Wiley & Sons. 1994. 590 p.
40 The Sadtler Standard Spectra: Infrared Grating Spectra. Vol 1-99. Philadelphia: Sadtler Research Laboratories. 1966-1990.
41 Искусственный интеллект. Применение в химии / Под ред. Е. Пирса, Б.М. Хони. М.: Мир. 1988. 430 с.
42 *Jurs P.C.* Computer Software Application in Chemistry. 2nd edition. New-York: Wiley & Sons. 1996. 291 p.
43 *Luinge H.J.* Automated Interpretation of Vibrational Spectra // Vib. Spectroscopy. 1990. V.1. P. 13-18.
44 *Эляшберг М.Е.* // Журн. аналит. химии. 1992. Т.47, №6. С. 966-981.
45 *Warr W.A.* // Anal. Chem. 1993. V.65. P. 1045A-1050A, 1087A-1095A.
46 *Эляшберг М.Е.* // Успехи химии. 1999. Т.68. С. 579-604.
47 *Подгорная М.И.,Derendяев Б.Г.* // НТИ. Сер.2. 1995, № 9. С.1-5.
48 *Averil D.F., Baird K.S., Hopkins L.L., Herkes M.J.* // J. Chem. Inf. Comput. Sci. 1990. V.30. P.133-136
49 *Koptyug V.A., Ulyanov G.P., Derendyaev B.G. e.o.* // CODATA Bull. 1981, V.40. P.45-53.
50 *Yamamoto O., Someno K., Wasada N. e.o.* // Anal. Sciences. 1988, V. 4. P.233-239.
51 *Grasselli J.G.* // Pure & Appl. Chem. 1987, V. 59, №5. P.673-681.
52 *McDonald R.S., Wilks P.F.* // Appl. Spectrosc. 1988. V. 42, №1. P.151-162.
53 *Davies A.M.* // Spectrosc. Int. 1991. V.3, №2. P.16-18.
54 *Sperline R.P.* // Appl. Spectrosc. 1991. V.45. P.1046-1047.
55 *Heller S.R.* // Chem. Int. 1985. V.13, № 6. P.224-231.
56 *Rumble Jr. J.R., Lide Jr. D.R.* // J. Chem. Inf. Comput. Sci. 1985. V.25, № 3. P.231-235.
57 *Heller S.R. //* Pure and Appl. Chem. 1995. V. 67. P.1027-1030.
58 Указатель литературных данных по спектроскопии ядерного магнитного резонанса ЯМР-^{1}Н / Под ред. В.А. Коптюга, М.И. Подгорной. Вып. 1-26. Новосибирск: НИОХ СО АН СССР. 1976-1988.
59 The Sadtler Standard Spectra. Nuclear Magnetic Resonance Spectra, V.1-94. Philadelphia: Sadtler Research Lab. 1966-1991.

60 *Pouchert C.J., Behnke J.* The Aldrich Library of C-13 and H-1 FT-Spectra. V.1-3. Milwaukee Aldrich Comp. 1992.

61 Handbook of Proton-NMR Spectra and Data. Asahi Research Center / Ed. S-I. Sasaki, V.1-10. Academic Press. 1986.

62 The Sadtler Standard Carbon-13 Nuclear Magnetic Resonance Spectra, V. 1-160. Philadelphia: Sadtler Research Lab. 1976-1991.

63 *Bremser W., Hardt A., Ernst L.* Carbon-13 NMR Spectral Data. Weinheim,N.Y.: VCH Verlagsgesellschart. 1987.

64 *Grasselli J.G., Ritchey W.M.* Atlas of Spectral Data and Physical Constants for Organic Compounds, 2nd Ed. V.1-6. CRC Press, 1975.

65 Атлас спектров углеродного магнитного резонанса. / Под ред. В.А.Коптюга. Вып.1-8. Новосибирск: НИОХ СО АН СССР. 1981-1990.

66 *Самитов Ю.Ю.* Атлас спектров ЯМР пространственных изомеров. Казань: изд-во КГУ. Т. 1 (1978). Т. 2. (1983).

67 *Atta-ur-Rahman, Ahmad V.U.* C13-NMR of Natural products. V.1, 2.. N-Y., London: Plenum. 1992.

68 *Derendyaev*

68 *Дерендяев Б.Г., Подгорная М.И., Осташевская Л.А. //* НТИ. Сер.2. 1995, № 1. С. 11-17.

69 C-13 NMR Search Libraries an IBM-PC // Appl. Spectrosc. 1988. V.42, № 3.

70 *Weast R.A., Astle M.J.* Handbook of Data Organic Compounds. CRC Press, 1985, V. 1-2.

71 *Heller S.R. //* J. Chem. Inf. Comput. Sci., 1985, V.25. P.224-231.

72 *Warr W.A.* // Chemomet. and Intel. Lab. Systems. 1991. V. 10. P.279-292.

73 *Yamamoto O., Someno K., Wasada N., a.o.* // Anal. Sci. 1988. V. 4. P.233-239.

74 *Pretsch E., Furst A., Badertscher M.a.o.* // J. Chem. Inf. Comput. Sci. 1992. V.32. P.291-295.

75 *Haas R., Strasser G., Scsibrang H. a.o.* // IV-Software-Entwickl. Chem 3. 1988. P.37-43.

76 *Speck D.D., Venkataraghavan R., McLafferty F.W.* // Org. Mass Spectrom. 1978. V.16. P.209-213.

77 *Yamamoto O., Yayamizu K., Yanagisawa M.* // Anal. Sci. 1988. V.4. P. 347-352.

78 *Zupan J.* // Anal.Chim.Acta. 1978. V.103, № 4. P.273-288.

79 *Codding E.G., Horlick G.* // Appl.Spectroscopy. 1973. V.27, № 5. P.366-370.

80 *Miller T.G. , Faulkner L.R.* // Anal.Chem. 1976. V.40, № 14. P.2083-2088.

81 *Pretsch E., Clerc J., Bendi J.* // Fresenius Z. Anal.Chem. 1986. Bd.324, № 7. S.344-349.

82 *Jim K., Miller T., Faulkner L.R.* // Anal. Chem. 1977. V.49. № 13. P.2069-2074.

83 *Mulkerin M.G., Wampler J.E.* // Anal.Chem. 1982. V.54. №11. P.1778-1782.

84 *Дерендяев Б.Г., Машков В.Е., Пиоттух-Пелецкий В.Н., Нехорошев С.А.* // Журн. структ. химии. 2001. Т.42, № 2. С.313-324.

85 *Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Шарапова О.Н.* // Журн. структ. химии. 2000, Т.41. С.378-389.

86 *Yamamoto O., Yayamizu K., Yanagisawa M.* // Anal. Sci. 1988. V.4. P.461-466.

87 *Henneberg D.* Adv. Mass Spectrom./ Ed by A. Quayle. London.: Academic Press. 1980, V.8B. P.1511-1515.

88 *Kwiatkowski J., Riepe W.* Adv. Mass Spectrom. / Ed by A. Quayle. London.: Academic Press. 1980. V.8B. P.1582-1585.

89 *Hollos J.* // Int. J. Mass Spectrom. Ion Phys. 1983. V. 47. P.321-324.

90 *Stauffer D.B., McLafferty F.W., Ellis R.D. a.o.* // Anal. Chem., 1985, V. 57, № 6, P. 1056-1060.

91 *Chapman J.R.* Computers in Mass Spectrometry. London.: Academic Press. 1978. 265 p.

92 *Pesyna S., McLafferty F.W.* in Determination of Organic Structures by Phesical Methods / Ed. by F.C. Nachod, J.J. Zuckerman, F.W. Randall. N-Y.: Academic Press. 1976. V.6. Chapt.2.

93 *Chapman J.R.* // J. Phys. E: Sci Instrum. 1980. V.13. P.365-375.

94 *Вершинин В.И.,Кайзер Л.В.,Негодова В.В. и др.* // Журн. прикл. спектроскопии. 1990. Т.53, №1. С.110-115

95 *Исидоров В.А., Зенкевич И.Г.* Хромато-масс-спектрометрическое определение следов органических веществ в атмосфере. Л.: Химия. 1982. 136 с.]

96 *Киршанский С.П., Лебедев К.С., Дерендяев Б.Г., С.А.Нехорошев* / Журн. аналит. химии. 1987. Т. 42, № 7. С.1320-1330.

97 *Sokolov S., Karnofsky J., Gustafson P.* // Finnigan Appl. Report. 1978, № 2, 45 p.

98 *Покровский Л.М., Дерендяев Б.Г.* // Журн. аналит. химии. 1988. Т.43, № 5. С.786-792.

99 *Buser Y.R., Arn H., Guerin P., Rauscher S.* // Anal. Chem. 1983. V. 55. P.818-822.

100 *Dromey R.G.* // Anal. Chem. 1979. V.51. P.229-232.

101 *Артемьев Б.В., Розынов Б.В., Михеева О.С.* // Журн. аналит. химии. 1980. Т. 35. С. 335-341.

102 *Cleij P., Van Klooster Y.A., Van Houwelinger J.C.* // Anal. Chim. Acta. 1983. V.150. P.23-36.

103 *Heller S.R., Budde W.L., Martinsen D.P., Milne G.W.A.* // Int. J. Mass Spectrom. 1983. V.18. P.ii.

104 *Pesyna G.M., Venkataraghavan R., Dayringer H.E., McLafferty F.W.* // Anal. Chem. 1976. V.48. P.1362-1368.

105 *Harrington P.B., Isenhour T.L.* // Anal. Chim. Acta. 1987. V.197. P.1298-1301.

106 *Clerc J.T., Pretch E., Zurcher M.* // Mikrochim. Acta. 1986. V.11. P.217-221.

107 *Stadalius M.A., Gold H.S.* // Anal.Chem. 1983. V.55. № 1. P.49.

108 *Wrabetz K.* // Fresenius Z. anal.Chem. 1976. Bd.272. №2. S.135.

109 Obukowicz J., Hippe Z. // J. Mol. Struct., 1986, V.142, P.17-20.

110 *McLafferty F.W., Venkataraghavan R., Kwok K.S., Pesyna G.* / Adv. Mass Spectrom. Ed. by A.R. West. London. 1974. P.999-1010.

111 *Delaney M.F., Hallowell, Jr. J.R., Warren V.F.* // J. Chem. Inf. Comput. Sci. 1985. V.25. P.27-30.

112 *Grotch S.L.* // Anal. Chem. 1970. V.42, P.1214-1222.

113 *Вершинин В.И.* В кн. "Применение математических методов и ЭВМ в аналитической химии". Проблемы аналитической химии. Т.9. М.: Наука, 1989. С.123-130.

114 *Henneberg D., Weimann B.* // Spectra (Finnigan MAT). 1984. V.10, P.11-19.

115 *Varmuza K., Werther W., Henneberg D., Weimann B.* // Rapid Comm. in Mass Spectrom. 1990. V.4, № 5. P.159-162.

116 *Piottukh-Peletsky V.N., Derendyaev B.G.* // Anal. Chim. Acta. 1999. V.396, P.99-103.

117 *Lowry S.R., Huppler D.A., Anderson C.R.* // J. Chem. Inf. Comput. Sci. 1985. V.25, P.235-241.

118 *Saeki S., Tanabe K.* // Appl. Spectrosc. 1984. V.38, P.693-697.

119 *Powell L.A., Hieftje G.M.* // Anal. Chim. Acta. 1978. V.100. P.313-320.

120 *Ehrentreich F.* // Fresenius J. Anal. Chem. 1997. V.359. P.56-60.

121 *Jung-Pin Y., Friedrich H.B.* // Appl. Spectrosc. 1987. V.41. P.869-874.

122 *Fuller M., Rosental R.* // SPIE-Int. Soc. Opt. Eng. 1993. V.2089. P.440-441.

123 *Penchev P.N., Sohou A.N., Andreev G.N.* // Spectrosc. Lett. 1996. V.29 P.1513-1522.

124 Ravak Y., Esen R. // J. Chem. Inf. Comput. Sci. 1993. V.33, P.595-597.

125 *Varmuza K., Penchev P.N., Scsibrany H.* // J. Chem. Inf. Comput. Sci. 1998. V.38. P.420-427.

126 *Лебедев К.С., Шарапова О.Н., Коробейничева И.К., Кохов В.А.* // Сиб. химический журн. 1993. Т.1, № 1. С.50-56.

127 *Sherman J.W., de Haseth J.A., Cameron D.G.* // Appl. Spectrosc. 1989. V.43. P.1311-1316.

128 *Дробышев Ю.П., Нигматуллин Р.С., Лобанов В.И. и др.* // Вестник АН СССР. 1970. Т.8. С.75-83.

129 *Kawata S., Noda T., Minami S.* // Appl. Spectrosc. 1987. V.41, P.1176-1188.

130 *Harrington P.B., Isenhour T.L.* // Appl. Spectrosc. 1987. V. 41. P.449-453.

131 *Harrington P.B., Isenhour T.L.* // Anal. Chem. 1988. V.60, P.2667-2670.

132 *Anderegg R.J., Pyo D.* // Anal. Chem. 1987. V.59. P.1914-1919.

133 *Zupan J., Munk M.E.* // Anal. Chem. 1985. V.57. P.1609-1615.

134 *Zupan J., Munk M.E.* // Anal. Chem. 1986. V.58. P.3219-3225.

135 *Cooper J.R., Wilkins C.L.* // Anal. Chem. 1988. V.61. P.1571-1576.

136 *Djerga J.M., Small G.W.* // Anal. Chem. 1990. V.62. P.226-233.

137 *Вершинин В.И., Топчий В.А.,Наумов С.Е.* // Журн. аналит. химии. 1987. Т.42, №.5. С.837-845.

138 *Киршанский С.П., Лебедев К.С., Дерендяев Б.Г.* // Изв. СО АН СССР. Сер. хим. наук. 1984. вып. 1. С. 97-103.

139 *Дерендяев Б.Г., Сухарев Ю.И., Хоц М.С.* Автоматизированные методы анализа масс-спектрометрической информации. Уфа. 1986. 119 с.

140 *Van Marlen G., Van`t Klooster H.A.* // Anal. Chem. 1979. V.51. P.420-423.

141 *Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Корнакова Т.А.* // Химия в интерес. устойч. развития. 2001. Т.9, № 1. С.17-26.

142 *Stein E.* // Pittsburgh Conf., Anal.Chem. and Appl.Spectrosc. Chicago. 1994. P.674.

143 *McLafferty F.W.* // Anal. Chem. 1979. V.51. P.1441-1443.

144 *Piottukh-Peletsky V.N., Derendyaev B.G.* // Anal. Chim. Acta. 2000. V.409. P.181-195.

145 *Hallower, Jr J.R., Delaney M.F.* // Anal. Chem. 1987. V.55. P.1544-1549.

146 *Harrington P.B., Isenhour T.L.* // Anal. Chim. Acta. 1987. V.197, P.1298-1301.

147 *Harrington P.B., Isenhour T.L.* // Appl. Spectrosc. 1987. V.41. P.146-161.

148 *Нехорошев С.А., Дерендяев Б.Г., Киршанский С.П., Лебедев К.С.* // Журн. аналит. химии. 1987. Т.42. № 7. С.1312-1320.

149 *Киршанский С.П., Дерендяев Б.Г.* // Журн. аналит. химии. 1997. Т.52. С.826-830.

150 *Покровский Л.М., Дерендяев Б.Г.* // Журн. аналит. химии. 1990. Т.45. С.2405-2411.

151 *Clerc J-T.* // Comp.-Enhanced Anal. Spectrosc. 1987. V.1. P.146-161.

152 *Rasmussen G.T., Isenhour T.L.* // Appl. Spectrosc. 1979. V.33. P.371-376.

153 *Ruprecht M., Clerc J-T.* // J. Chem. Inf. Comput. Sci.. 1985. V.25. P.241-244.

154 *Affolter C., Clerc J-T.* // Fresenius J. Anal. Chem. 1992. V.344, №4-5. P.136-139.

155 *Лебедев К.С., Дерендяев Б.Г.* // Химия в интерес. устойч. развития. 1995, Т.3, С.269-285.

156 *Зорин Б.Я., Волкович С.В., Гришин Н.Н. и др.* // Химия древесины. 1987, №5. С.102-106.

157 Vereshchagin S.N., Kirik N.P.,. Shishkina N.N, a.o.. *Catalysis Today.* 2000. V.61. P.129-136.

158 *Lindsay R.K., Buchanan B.G., Feigenbaum E.A. and Lederberg J.* Applications of Artificial Intelligence for Chemical Inference: the DENDRAL Project. N.Y.: McGraw-Hill, 1980.

159 *Funatsu K., Sasaki S.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.190-204.

160 *Эляшберг М.Е., Грибов Л.А., Колдашев В.Н., Плетнев И.В.* // Докл. АН СССР. 1983. Т.268. С.112- 119.

161 *Elyashberg M.E.,.Martirosian E.R., Karasev Y.Z., Thile H.* // Anal. Chem. Acta. 1997. V.337. P.265-285.

162 *Bremser W., Grzonka M.* // Mikrochim. Acta. 1991. V.2. P.483-491.

163 *Will M., Fachinger W., Richert J.R.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.221-227.

164 *Коптюг В.А., Бочкарев В.С., Дерендяев Б.Г. и др.* // Журн. структур. химии. 1977. Т.18. С.440-459.

165 *Lebedev K.S., Tormyshev V.M., Derendyaev B.G., Koptyug V.A.* // Anal. Chim. Acta. 1981. V.133. P.517-525.

166 *Лебедев К.С., Отмахова Е.А., Гриценко И.В.* // Изв. СО АН СССР. Сер. хим. наук. 1990. №5. С.73-79.

167 *Лебедев К.С.* // Журн. аналит. химии. 1993. Т. 48. С.851-863.

168 *Дерендяев Б.Г., Нехорошев С.А., Лебедев К.С., Киршанский С.П.* // Журн. аналит. химии. 1991. Т.46. С.1870-1879.

169 *Derendjaev B.G., Nekhoroshev S.A., Lebedev K.S., Kirshansky S.P.* // J. Chem. Inf. Comput. Sci. 1992. V.32. P.255-260.

170 *Киршанский С.П., Дерендяев Б.Г.* // Журн. аналит. химии. 1997. Т.52. С.826-830.

171 *Fuchs Ph.L. and Bunnell Ch.A.* // Carbon-13 NMR Based Organic Spectral Problems. N.Y.: Wiley, 1979.

172 *Лебедев К.С., Киршанский С.П.* // Изв.СО АН СССР. Сер. хим. наук. 1989, № 4. С.79-87.

173 *Лебедев К.С., Нехорошев С.А., Киршанский С.П., Дерендяев Б.Г.* // Сиб. хим. журнал. 1992. № 3. С.72-79.

174 *Derendyaev B.G., Lebedev K.S., Nekhoroshev S.A., Kirshansky S.P.* // Comput. Chem. 1994. V.18. P.81-88.

175 *Varmuza K.* Pattern Recognition in Chemistry. Berlin: Springer-Verlag, 1980.

176 *Hippe Z.H.* // Chem. Anal. 1995. V.40. P.473-479.

177 *Zupan J., Gasteiger J.* Neural Networks for Chemists. Weinheim: VCH, 1993.

178 *Munk M.E., Madison M.S., Robb E.W.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.231-238.

179 *Klawun C., Wilkins C.L.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.249-257.

180 *Eghbaldar A., Forrest T.P., Cabrol-Bass D., a.o.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.637-643.

181 *Haraki R.S., Venkataraghavan R.., McLafferty F.W.* // Anal. Chem. 1981. V.53. P.386-392.

182 *Bremser W. , Fachinger W.* // Magn. Reson. Chem. 1985. V.23. P.1056-1071.

183 *Bremser W., Neudert R.* // Eur. Spectrosc. News. 1987. № 75. P.10-27.

184 *Stein S.E.* // J.Am.Soc.Mass Spectrom. 1995. V.6. P.644-655.

185 *Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Богданова Т.Ф.* //Журн. структ. химии. 1997. Т.38. С.370-379.

186 *Лебедев К.С., Тормышев В.М., Шарапова О.Н., Мамаева Н.В., Дерендяев Б.Г., Коптюг В.А.* // Изв. СО АН СССР. Сер.хим.наук. 1980, №.2. С.54-64.

187 *Лебедев К.С., Тормышев В.М., Дерендяев Б.Г.* // Изв. СО АН СССР. Сер. хим. наук. 1980, №.2. С.64-73.

188 *Cone M.M., Venkataraghavan R., McLafferty F.W.* // J.Amer.Chem.Soc. 1977, V.99, P.7688-

7671.
189 *Лебедев К.С., Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Коптюг В.А.* // Изв. СО АН СССР. Сер. хим. наук. 1982. №1. С.105-114.
190 *Dayringer Y.E., Pesyna C.M., Venkataraghavan R. and McLafferty F.W.* // Org. Mass. Spectrom. 1976. V.11, P.529.
191 *Лебедев К.С., Строков И.И., Поддубный К.С., Кохов В.А.* // Тез. докл. IX Всесоюз. Конф. "Химическая информатика", Черноголовка: ИФАВ РАН, 1992, С.41.
192 *Lebedev K.S. , Cabrol-Bass D.* // J. Chem. Inf. Comput. Sci. 1998, V.38, P.410-419.
193 *Лебедев К.С., Гриценко И.В.* // Журн. структ. химии. 1993. Т.34. С.51-60.
194 *Строков И.И.., Гриценко И.В., Лебедев К.С.* // Изв. СО АН СССР. Сер. хим. наук. 1987. №2. С.78-85.
195 *Carhart R.E., Smith D.H., Gray N.A.B., Nourse J.G., Djerassi C.* // J. Org. Chem. 1981. V.46. P.1708.
196 *Funatsu F., Miyabayashi N., Sasaki S.* // J. Chem. Inf. Comput. Sci. 1988. V.28. P.18-28.
197 *Christie B.D. , Munk M.E.* // J. Chem. Inf. Comput. Sci. 1988. V.28. P.87-93.
198 *Wieland T., Kerber A., Laue R.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.413.
199 *Молодцов С.Г., Лебедев К.С., Дерендяев Б.Г.* // Журн. структур. химии. 1994. Т.35. С.46-
200 *Киршанский С.П., Молодцов С.Г., Лебедев К.С.* // Изв. СО АН СССР. Сер. хим. наук. 1989, №.5. С.3-10.
201 *Molodsov S.G.* // Commun. Math. Chem. (MATCH). 1994, N30. P.213-224.
202 *Molodsov S.G.* // Commun. Math. Chem. (MATCH). 1998, V.37. P.157-.
203 *Дерендяев Б.Г., Лебедев К.С., Строков И.И. и др.* // Химия в интересах устойч. развития. 1998, Т.6. С.25-39.
204 *Strokov I.I., Lebedev K.S.* // J. Chem. Inf. Comput. Sci. 1999. V.39. P.659-665
205 *Furst A., Pretsch E.* // Anal. Chem. Acta. 1990. V.229. P.17-
206 *Bremser W.* // Magn.Reson.Chem. 1985. V.23. P.271-
207 *Schweitzer R.C., Small G.W.* // J. Chem. Inf. Comput. Sci. 1997. V.37. P.249-
208 *Burgin S.R., Munk M.E., Pretsch E.* // J. Chem. Inf. Comput. Sci. 1996. V.36. P.239-
209 *Gasteiger J., Hanebeck W., Schulz K.-P.* // J. Chem. Inf. Comput. Sci. 1992. V.32, P.264-271.
210 *Gribov L.A., Zinovyev K.A.* // J. Mol. Struct. 1992. V.268. P.191-
211 *Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Молодцов С.Г., Богданова Т.Ф.* // Журн. структ. химии. 1997. Т.38. С.786-794.
212 *Dubois J.E., Mathieu G., Peguet P. a.o.* // J. Chem. Inf. Comput. Sci. 1990. V.30. P.290-.
213 *Chen L., Robien W.* // Anal. Chim. Acta. 1993. V.272. P.301-308.
214 *Small G.W.* // J. Chem. Comput. Sci. 1992. V.32. P.279-285.

215 *Baumann K., Clerc J.T.* // Anal. Chem. Acta. 1997. V.348. P.327-343.

216 *Kalchhauser H., Robien W.* // Chem. Inf. Comput. Sci. 1985. V.25. P.103-108.

217 *Passlack M., Bremser W.* “IDIOTS – Structure-oriented Databank System for the Identification and Interpretation of Infrared Spectra”. In *Computer-Supported Spectroscopic Databases* / J.Zupan (Editor). Chichester: Ellis Horwood. 1986. P.92-117.

218 *Cadish M., Munk M.E., Clerc J.T., Pretshc E.* // Chem. Inf. Comput. Sci. 1992. V.32. P.286-290.

219 *Barth A.* // Chem. Inf. Comput. Sci. 1993. V.33. P.52-58.

220 *Debska B.J.* // Comput. Chem. 1995. V.19. P.269-275.

221 *Neudert R., Penk M.* // Chem. Inf. Comput. Sci. 1996. V.36. P.244-248.

222 *Киршанский С.П., Лебедев К.С., Дерендяев Б.Г.* // Журн. аналит. химии. 1987. Т.42. С.1092-1097.

223 *Strokov I.I., Lebedev K.S.* // Chem. Inf. Comput. Sci. 1996. V.36. P.741-745.

224 *Лебедев К.С., Строков И.И., Гриценко И.В. и др.* // Тез. докл. VI Всеросс. Конф. “Аналитика Сибири и Дальнего Востока – 2000”, Новосибирск. 2000. С.129-130.

225 *Strokov I.I.* // J. Chem. Inf. Comput. Sci. 1995. V.35. P.939-944.

226 *Строков И.И., Лебедев К.С., Дерендяев Б.Г.* // Журн. структ. химии. 1996. Т.37. С.1129-1139.

227 *Столяров Б.В., Савинов И.М., Витенберг А.Г.* Руководство к практическим работам по газовой хроматографии. Л.: Химия. 1988. С.179.

228 *Зенкевич И.Г., Цибульская И.А.* // Журн.аналит.химии. 1989. Т.44, № 1. С.90-96.

229 *Ударов Б.Г., Куликов В.И.* // Исследование хроматографических процессов. М.: НИИТЭХим. 1982. С.35.

230 *Другов Ю.С.* Экологическая аналитическая химия. М.: Изд-во “Анатолия”. 2000. С.48-64.

231 *Другов Ю.С., Родин А.А.* Газохроматографическая идентификация загрязнений воздуха, воды и почвы. Практическое руководство. СПб.: Теза. 1999. 622 с.

232 *Гоишон Ж., Гийемен К.* Количественная газовая хроматография для лабораторных анализов и промышленного контроля. М., Мир. 1991. Ч.II. С.5-16.

233 *Вигдергауз М.С., Курбатова С.В.* Организация ЭВМ-банка справочных данных по хроматографическому удерживанию органических соединений и использование его для анализа объектов окружающей среды. Горький: Горьковский госуниверситет, 1990. С.3-10.

234 *McReynolds W.O.* Gas Chromatography Retention Data. Preston Technical Abstract. Evanston (USA), 1966.

235 *Полухин Д.Ю., Ревельский И.А., Яшин Ю.С.* // Зав. лаборатория. 1999. Т.65, № 3. С.3-6.

236 *Вигдергауз М.С., Колосова Е.А., Курбатова С.В.* // Зав. лаборатория. 1993. №6. С.7-10.

237 *Григорьева Д.Н., Головня Р.В.* // Журн.аналит.химии. 1985. Т.40, № 10. С.1733.

238 *Вигдергауз М.С., Арутюнов Ю.И., Курбатова С.В., Колосова Е.А.*// Журн. аналит.химии. 1994. Т.49, № 10. С.1067-1072.

239 *Schupp O.E., Lewis J.S.* Compilation of Retention Data. ASTM Data Series. Publication DS-25A. Philadelphia: ASTM. 1967.

240 *Зенкевич И.Г.* // Журн.структурной химии. 1994. Т.35, № 6. С.176-182.

241 *Takacs J.M.* // J. Chromatogr. Sci. 1991. V.91. P.382.

242 *Зенкевич И.Г., Кузнецова Л.М.* // Докл. АН СССР. 1990. Т.315. № 4. С.881-885.

243 *Sato Y., Kira A., Horiba Ltd.* // Патент США N 134336 (приоритет от 11.10.92). МКИ G01 N 23/223. РЖХим 1996 24Г41П.

244 *Шатц В.Д., Авотс А.А., Кофман А.А., Силис Я.Я.* // Журн. аналит.химии. 1975. Т.30. № 12. С.2306-2310.

245 *Вигдергауз М.С., Курбатова С.В.*// Журн.аналит.химии.1991. Т.46. № 4. С.683-694.

246 Зенкевич И.Г., Васильев А.В. // Журн.аналит.химии. 1993. Т.48, № 3. С. 473-485.

247 Kuwata K, Yamada K. // J. Chromatogr. 1983. V.256. № 2. P.303-312.

248 *Boessencool H.J., Cleij P., Goewie C.E et al.* // Mikrochim. Acta. 1986. №II. p.75-92

249 *Driscoll J.N., Ferioli P., McDermott G. et al.* // [Pap.] Pittsburgh Conf. 1986. v.1. p.290. РЖХим 1986 - 22Г 446.

250 PIANO Manual. / Analytical Automation Instr., Inc (Italy). Ver. 6.15. 1993

251 *Below E., Burmann M.* // Liquid Chromatogr. 1994. V.17. № 20. *P*.4134-4144.

252 *Семенов А.Д., Сапожникова Е.В., Бондаренко Ю.Ю.* // Известия ВУЗов. Северо-Кавказский регион. Естественные науки. - 1998. № 4. С.92-95.

253 *Березкин В.Г., Герасименко В.А., Набивач В.М.* // Журн.аналит.химии. 1996. Т.51, №4. С.410-418.

254 *Вершинин В.И., Топчий В.А., Медведовская И.И.* // Журн. аналит. химии. 2001. Т.56, № 4. С.457-465.

255 *Вершинин В.И., Топчий В.А.* // Журн. аналит.химии. 1989. Т. 44, № 6. С.1085-1093.

256 *Шаевич А.Б., Мучник В.Л.* // Журн. аналит.химии. 1990. Т.45, № 8. С.1526-1531.

257 *Светлова Н.Н., Григорьева Д.Н., Журавлева И.Л., Головня Р.В.* // Журн. аналит.химии. 1984. Т.39, №7. С.1292-1296.

258 *Leathard D.A., Shurlock B.C.* Identification techniques in gas chromatography. N.Y.:Wiley Interscience. J.Wiley and Sons Ltd. 1970. P.44.

259 *Данцер К., Тан Э., Мольх Д.* Аналитика.Систематический обзор. М.: Химия.1981.С.44-52.

260 *McLafferty F.W.* // Anal.Chem. 1977. V.49, № 9. P.1442.

261 *Соколова О.В., Ильичева Н.Б., Медведовская И.И., Вершинин В.И.* // Аналитика и контроль.

2000, № 4. С. 460-475.

262 *Вигдергауз М.С., Петрова Е.Н., Шатских С.Я.* // Журн. аналит.химии. 1989. Т.44. № 4. С.712-720.

263 ASTM D 5134. Standard Test Method for Detailed Analysis of Petroleum Naphthas through n-Nonane by Capillary Chromatography.

264 *Теплицкая Т.А., Вершинин В.И., Овечкин А.Б.* // Журн. аналит. химии. 1982. Т.37, № 7. С.1256-1262.

265 *Вершинин В.И., Ильичева Н.Б., Медведовская И.И. и др.* // VI Конференция "Аналитика Сибири и Дальнего Востока – 2000" Тезисы докладов. Новосибирск, 2000, с.13.

266 *Сидоров Р.И., Хвостикова А.А., Вахрушева Г.И.* // Журн.аналит.химии. 1073. Т.28, № 8. С.1593-1597.

267 *Зенкевич И.Г.* // Журн.прикл.химии. 1994. Т.67, № 11. С.1877-1882.

268 *Топчий В.А., Дворкин П.Л.* // Новые информационные технологии в университетском образовании. Новосибирск. 1997. С.31-35.

269 *Топчий В.А., Вершинин В.И.* Омский научный вестник. 2000. № 1. С.39-41.

270 Лекомцев А.С., Назаркин А.В., Куклинский А.Я., Рябчук Г.В. // Экологическая химия. 1999. Т.8. № 2. С.80-87.

271 *Marquart R.G. , Katsnelson I.* // J. appl. Crystallogr. 1979. V.12, № 10. P.629-634.

272 *Бурова Е.М., Жидков Н.П., Зубенко В.В. и др.* // Докл. АН СССР. 1977. Т.232. N 5. С.1066-1068.

273 *Хоц М.С.* Масс-спектральный анализ многокомпонентных смесей органических соединений и его математическое обеспечение. Дисс...: докт.физ.-мат.наук. М. 1986. - 328 с.

274 *Хоц М.С.* // Математические методы и ЭВМ в аналитической химии. М.: Наука. 1989. С.87-103.

275 *Гречушников Б.Н.* // Математические методы и ЭВМ в аналитической химии М.: Наука. 1989.С.26-42.

276 *Волков В.В.* // Математические методы и ЭВМ в аналитической химии. М.: Наука. 1989. С. 42-50.

277 *Domokos L., Henneberg D.* // Anal.Chim.Acta. 1984. V.65, № 1. P.75-86.

278 *Хоц М.С., Белан Г.Б., Шапиро И.С.* // Журн. аналит.химии. 1985. Т.40, № 11. С.2052-2056.

279 *Туров Ю.П.* Применение масс-спектрометрии при исследовании состава смесей. Дисс...канд. физ.-мат. наук. Кемерово, 1980. -186 с.

280 *Безъязыкова А.Н., Смирнов Ю.Н., Кривошеева Л.В. и др.* / Гигиена и санитария. 1989.

N.11, с.68-71.

281 *Нахмансон М.С., Черный Ю.А.* // Зав. лаборатория. 1985. N 5. С.23-29.

282 *Вершинин В.И, Санина О.В.* // Ж. прикл. спектроскопии. 1988. Т.48. № 2. С.248-252.

283 *Дозморов С.В.,Шалаева М.Е.,Сизиков А.М.* // Ж. прикл. спектроскопии. 1984. Т.40. № 5. С.863-864.

284 *Пуговкин Н.М., Ольховик Г.А., Вылегжанин О.Н.* // Журн. прикл. спектроскопии. 1982. Т.36. N 5. С.766-770.

285 *Пиккеринг У.* Современная аналитическая химия. М.: Химия,1977. С.97.

286 *Мурашова В.И., Тананаева А.Н., Ховякова Р.Ф.* Качественный химический дробный анализ. М.: Химия, 1976. С.12.

287 *Кане И.А., Консон Е.Д., Нахмансон М.С. и др.* // Аппаратура и методы рентгеновского анализа. Вып.32. Л.: Машиностроение,1984,с.62-78.

288 *Вершинин В.И., Дозморов С.В., Овечкин А.Б.* // Ж.аналит.химии. 1985. Т.40. N 5. с. 2249-2258.

289 *Вершинин В.И., Дозморов С.В.* // Ж. прикл. спектроскопии. 1985. Т.42, N.6, с.944-948.

290 *Шараф. Ковальски.* Хемометрика.

291 *Алексеева Т.А., Теплицкая Т.А.* Спектрофлуориметрические методы анализа ароматических углеводородов в природных и техногенных средах. Л.: Гидрометеоиздат, 1981. 215 с.

292 *Теплицкая Т. А., Алексеева Т. А., Вальдман М. М.* Атлас квазилинейчатых спектров люминесценции ароматических молекул. М.: Изд-во МГУ, 1978. 22 с.

293 *Colmsjo A., Ostman C.* Atlas of Shpol'skii spectra and other low temperature spectra of POM. Stockholm. Univ. Stockholm, 120 p.

294 *Karcher W., Fordcham R.J., Dubois J.J. e.o.* Spectral Atlas of polycyclic compounds. Dordrecht, 1985. -800 p.

295 *Horlick G.* // Anal. Chem. 1973. V.45, N2. P.319-324.

296 *Чарыков А. К.* Математическая обработка результатов химического анализа. Л.: Химия, 1984. 168 с.

297 *Aarnio P.A., Ala-Heikkila J.J., Hakulinen T.T. et al.* // J. Radioanal. Nucl. Chem. Art. 1995. V.193. N 2. p.219-227.

298 *Stauffler D.B., McLafferty F.W., Ellis R.D. et al.* // Anal.Chem. 1985. V.57, № 6, P.1056-1060.

299 *Blaffert T.* // Anal. Chim. Acta. 1984. V.161. № 1. P.135-148.

300 *Нахмансон М.С., Черный Ю.А.* Система автоматического рентгенофазового анализа АРФА. М.: ВНИИ научного приборостроения. !981. 63с.

301 *Colmsjo A., Stenberg U.* // Anal. Chem. 1979. V.51, №1. P. 145-150.

302 *Смирнов Ю.Н.* Определение полициклических ароматических углеводородов в сточных водах нефтеперерабатывающих предприятий методов низкотемпературной люминесценции. Дисс....канд.хим.наук. Одесса. 1989 - 178 с.

303 *Вершинин В.И., Целищев В.А., Чиркова Е.А.* // Журн. аналит. химии. 1993. Т.48, N 5, с.919-920.

304 *Зайдель А.Н., Прокофьев В.К., Райский С.М. и др*. Таблицы спектральных линий. М. : Наука, 1969. 782 с.